SELF-SUPERVISED LEARNING AND ITS APPLICATIONS IN MEDICAL IMAGE ANALYSIS

SILADITTYA MANNA

COMPUTER VISION AND PATTERN RECOGNITION UNIT

INDIAN STATISTICAL INSTITUTE, KOLKATA

April 2025

SELF-SUPERVISED LEARNING AND ITS APPLICATIONS IN MEDICAL IMAGE ANALYSIS

A thesis submitted to the Indian Statistical Institute
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

by

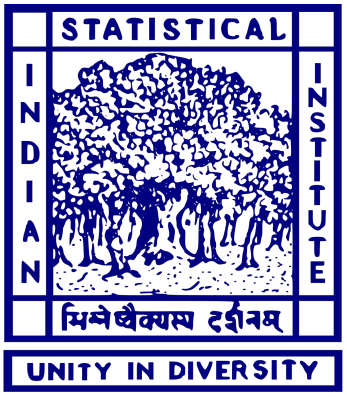
Siladittya Manna

Computer Vision and Pattern Recognition Unit
Indian Statistical Institue, Kolkata

Under the supervision of

Prof. Umapada Pal

Computer Vision and Pattern Recognition Unit
Indian Statistical Institue, Kolkata



Computer Vision and Pattern Recognition Unit
Indian Statistical Institute, Kolkata

April 2025

To everyone who stood beside me when I lost hope

Certificate

This is to certify that the thesis entitled **"SELF-SUPERVISED LEARNING AND ITS APPLICATIONS IN MEDICAL IMAGE ANALYSIS"**, submitted by **SILADITTYA MANNA** to the Indian Statistical Institute, Kolkata, for the award of the degree of **Doctor of Philosophy** in Computer Science, is a record of the original, bonafide research work carried out by him under my supervision and guidance. The thesis has reached the standards that meet the requirements of the regulations related to the award of the degree.

The results contained in this thesis have not been submitted in part or in full to any other University or Institute for the award of any degree or diploma to the best of my knowledge.

Umapada Pal

Prof. Umapada Pal
Computer Vision and Pattern Recognition Unit,
Indian Statistical Institute, Kolkata.

Date: 18/09/2024

Acknowledgements

My Ph.D. journey has been nothing short of a rollercoaster ride. There have been times when the acceptance of manuscripts prepared through untiring toil has overwhelmed my insignificant soul, as also when my confident self has been crushed by the unfortunate event of manuscript rejection. While we as humans often get people to celebrate with us in good times, only a few stand with us in dire times, when we seem to have lost hope.

Foremost, I would like to express my deepest gratitude to my supervisor, Prof. Umapada Pal, for providing me the opportunity to work with him and opening my mind to a plethora of possibilities. Additionally, I cannot be grateful enough for his guidance, invaluable support, freedom of thought, and insightful feedback throughout the journey of this thesis. His expertise, patience, and encouragement have been instrumental in shaping the direction and quality of this thesis, and also as a researcher.

Through this Ph.D. I was introduced to one of the nicest persons I ever met, my mentor, Dr. Saumik Bhattacharya. From studying new things and exploring new research ideas to trying out extracurriculars, there was no dearth of encouragement from my mentor. This allowed me to explore different ideas with an open mind and played an important role in shaping this thesis. I cannot show enough gratitude to Dr. Bhattacharya for his constant support, encouragement, and scrutiny of my work which have helped me improve as a researcher during my Ph.D.

I am also immensely thankful to the members of the Research Fellow Advisory Committee, Computer and Communication Sciences Division (RFAC-CCSD), for their valuable input, constructive criticism, and scholarly guidance. Their diverse perspectives have enriched this research and helped me navigate through its complexities.

I extend my heartfelt appreciation to the Indian Statistical Institute, Kolkata for providing the necessary resources and facilities essential for the completion of this thesis. The scholarly environment and academic resources available have significantly contributed to the thesis.

My sincere thanks go to my family members for their unconditional love, encouragement, and belief in me. They have been my source of strength and motivation throughout this academic journey.

Lastly, I am grateful to all my friends and colleagues who have supported me emotionally and intellectually during the ups and downs of this thesis.

This thesis would not have been possible without the support and contributions of all those mentioned above. While their names may appear on these pages, their impact resonates deeply in every aspect of this work.

Thank you.

Siladittya Manna 18/09/2024

Siladittya Manna

Abstract

Self-supervised learning (SSL) enables learning robust representations from unlabeled data and it consists of two stages: pretext and downstream. The representations learnt in the pretext task are transferred to the downstream task. Self-supervised learning has applications in various domains, such as computer vision tasks, natural language processing, speech and audio processing, etc. In transfer learning scenarios, due to differences in the data distribution of the source and the target data, the hierarchical co-adaptation of the representations is destroyed, and hence proper fine-tuning is required to achieve satisfactory performance. With self-supervised pre-training, it is possible to learn representations aligned with the target data distribution, thereby making it easier to fine-tune the parameters in the downstream task in the data-scarce medical image analysis domain.

The primary objective of this thesis is to propose self-supervised learning frameworks that deal with specific challenges. Initially, jigsaw puzzle-solving strategy-based frameworks are devised where a semi-parallel architecture is used to decouple the representations of patches of a slice from a magnetic resonance scan to prevent learning of low-level signals and to learn context-invariant representations. The literature shows that contrastive learning tasks are better than context-based tasks in learning representations. Thus, we propose a novel binary contrastive learning framework based on classifying a pair as positive or negative. We also investigate the ability of self-supervised pre-training to boost the quality of transferable representations. To effectively control the uniformity-alignment trade-off, we re-formulate the binary contrastive framework from a variational perspective. We further improve this vanilla formulation by eliminating positive-positive repulsion and amplifying negative-negative repulsion. The reformulated binary contrastive learning framework outperforms the state-of-the-art contrastive and non-contrastive frameworks on benchmark datasets. Empirically, we observe that the temperature hyper-parameter plays a significant role in controlling the uniformity-alignment trade-off, consequently determining the downstream performance. Hence, we derive a form of the temperature function by solving a first-order differential equation obtained from the gradient of the InfoNCE loss with respect to the cosine similarity of a negative pair. This enables controlling the uniformity-alignment trade-off by computing an optimal temperature for each sample pair. From experimental evidence, we observe that the proposed temperature function improves the performance of a weak baseline framework to outperform the state-of-the-art contrastive and non-contrastive frameworks. Finally, to maximise the transferability of representations, we propose a self-supervised few-shot segmentation pretext task to minimise the disparity between the pretext and downstream tasks. Using the Felzenszwalb-based segmentation method to generate the pseudo-masks, we train a segmentation network that learns representations aligned with the downstream task of one-shot segmentation. We propose a correlation-weighted prototype aggregation step to incorporate contextual information efficiently. In the downstream task, we conduct inference without fine-tuning and the proposed self-supervised one-shot framework performs better or at par with the contemporary self-supervised segmentation frameworks.

In conclusion, the proposed self-supervised learning frameworks offer significant improvements in representation learning, and enhancing performance on downstream medical image analysis tasks, as observed from the different experimental results of the thesis.

List of Publications

Journals

1. Siladittya Manna, Saumik Bhattacharya and Umapada Pal, "Self-Supervised Representation Learning for Detection of ACL Tear Injury in Knee MR Videos", Pattern Recognition Letters, Vol. 154, Pages 37-43, 2022 (DOI: 10.1016/j.patrec.2022.01.008).

2. Siladittya Manna, Saumik Bhattacharya and Umapada Pal, "Self-Supervised Representation Learning for Knee Injury Diagnosis from Medical Data", IEEE Transactions on Artificial Intelligence, Volume 5, Issue 4, Pages 1613-1623, 2024 (DOI: 10.1109/TAI.2023.3299883).

3. Siladittya Manna, Saumik Bhattacharya and Umapada Pal, "Self-Supervised Visual Representation Learning for Medical Image Analysis: A Comprehensive Survey", Transactions on Machine Learning Research, Vol. 2024,
URL: https://openreview.net/forum?id=3Wg1oErMcJ (Accepted).

4. Siladittya Manna, Saumik Bhattacharya and Umapada Pal, "MIO: Mutual Information Optimization using Self-Supervised Binary Contrastive Learning", ArXiv: abs/2111. 12664/, Submitted after major revision at IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI-2024-02-0270).

5. Siladittya Manna, Soumitri Chattopadhyay, Rakesh Dey, Saumik Bhattacharya and Umapada Pal, "Dynamically Scaled Temperature in Self-Supervised Contrastive Learning", ArXiv: abs/2308.01140/, Submitted after major revision at IEEE Transactions on Artificial Intelligence (TAI-2024-Jun-A-00829).

Presentations and proceedings in International/National Conferences

1. Siladittya Manna, Saumik Bhattacharya and Umapada Pal, "Interpretive Self-Supervised Pre-training: Boosting Performance on Visual Medical Data", Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'21), Pages 1–9, 2021 (DOI: 10.1145/3490035.3490273).

2. Siladittya Manna, Saumik Bhattacharya and Umapada Pal, "Correlation Weighted Prototype-based Self-Supervised One-Shot Segmentation of Medical Images", 27th International Conference on Pattern Recognition (ICPR), 2024 (Accepted).

Other Publications (Not included in this thesis)

1. Siladittya Manna, Dipayan Das, Saumik Bhattacharya, Umapada Pal and Sukalpa Chanda, "PLSM: A Parallelized Liquid State Machine for Unintentional Action Detection", IEEE Transactions on Emerging Topics in Computing, vol. 11, no. 2, pp. 474-484, 2023 (DOI: 10.1109/TETC.2022.3211011).

2. Siladittya Manna, Soumitri Chattopadhyay, Saumik Bhattacharya and Umapada Pal, "SWIS: Self-Supervised Representation Learning for Writer Independent Offline Signature Verification", IEEE International Conference on Image Processing (ICIP). pp. 1411-1415, 2022 (DOI: 10.1109/ICPR56361.2022.9956442).

3. Soumitri Chattopadhyay, Siladittya Manna, Saumik Bhattacharya and Umapada Pal, "SURDS: Self-Supervised Attention-Guided Reconstruction and Dual Triplet Loss for Writer Independent Offline Signature Verification", 26th International Conference on Pattern Recognition (ICPR). pp. 1600-1606, 2022 (DOI: 10.1109/ICPR56361.2022.9956442).

4. Subhajit Maity, Sanket Biswas, Siladittya Manna, Ayan Banerjee, Josep Lladós, Saumik Bhattacharya and Umapada Pal, "SelfDocSeg: A Self-supervised Vision-Based Approach Towards Document Segmentation", Proceedings of the 17th International Conference on Document Analysis and Recognition, Part I, pp. 342–360, 2023 (DOI: 10.1007/978-3-031-41676-7_20).

Contents

List of Figures

List of Tables

Abbreviations

Acronym	**W**hat (it) **S**tands **F**or
SSL	Self-Supervised Learning
MNIST	Modified National Institute of Standards and Technology
SVM	Support Vector Machine
PCA	Principle Component Analysis
tSNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
HoG	Histogram of Gradients
SIFT	Scale Invariant Feature Transform
SURF	Speeded Up Robust Features
ISOMAP	Isometric Mapping
LDA	Linear Discriminant Analysis
VAE	Variational Auto Encoder
NLP	Natural Language Processing
CV	Computer Vision
MRI	Magnetic Resonance Imaging
MS-COCO	Microsoft Common Objects in Context
CT	Computed Tomography
CNN	Convolutional Neural Networks
GAN	Generative Adversarial Networks
VQ	Vector Quantized
NCE	Noise Contrastive Estimation
HSIC	Hilbert-Schmidt Independence Criterion
MSE	Mean Squared Error
ResNet	Residual Network
ViT	Vision Transformer
CIE	Commission internationale de l'éclairage
GradCAM	Gradient Class Activation Mapping
AUC	Area Under Curve
LSTM	Long-Short Term Memory

ConvLSTM	Convolutional Long-Short Term Memory
ReLU	Rectified Linear Unit
SCUMBLE	Score of ConcUrrence among iMBalanced LabEls
LARS	Layer-wise Adaptive Rate Scaling
ABN	Abnormality
ACL	Anterior Cruciate Ligament
MEN	Meniscus
CI	Confidence Interval
MLP	Multiple Layer Perceptron
MLE	Maximum Likelihood Estimation
CIFAR	Canadian Institute for Advanced Research
STL	Self-Taught Learning
SGD	Stochastic Gradient Descent
ISIC	International Standard Industrial Classification of All Economic Activities
MHIST	Minimalist Histopathology Image Analysis Dataset
MURA	Musculoskeletal Radiographs
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
HN	Hard Negative
AMP	Automatic Mixed Precision
BN	Batch Normalization
kNN	k-Nearest Neighbour
SMvXX	Shifted Minima version XX
FSS	Few Shot Segmentation
MICCAI	Medical Image Computing and Computer Assisted Intervention Society
SPIR	Spectral Presaturation with Inversion Recovery
LLM	Large Language Models

Symbols

$\aleph$	Number of samples in the dataset
N	Batch size
H	height of input image
W	width of input image
C	Number of classes in the pretext or downstream task
$\mathcal{F}$	A slice of an MR scan
$\mathcal{U}$	Uniform distribution
$\mathcal{T}$	Augmentation or arrangement to be applied on an image
N_C	Number of classifiers
T_p	Number of positive pairs
T_n	Number of negative pairs
τ	Temperature hyperparameter
x	A sample or image
$\mathbb{R}^n$	n-dimensional Real Space
χ	Input Pair Space
X	Sampled Input Pair from χ
$\mathcal{X}_+$	Set of all positive pairs
$\mathcal{X}_-$	Set of all negative pairs
$f(\cdot)$	Output of network f
θ	Encoder Parameters
f_θ	Encoder f with parameters θ
h / h_θ	Feature vector obtained after passing x through f_θ.
ψ	Projector Parameters
g_ψ	Projector g with parameters ψ
z	Feature vector obtained after passing x through f_θ and g_ψ.
$\mathcal{Z}$	Space of projected latent feature vectors
c_{ij}	Cosine similarity between feature vectors of samples in a pair (x_i, x_j).
$s(z_i, z_j)$	Scoring function, calculate a pre-defined metric score between z_i and z_j.
p_+	Distribution of positive pairs

p_-	Distribution of negative pairs
$P_+^{i,j}$	Probability of obtaining the positive pair (x_i, x_j) or (z_i, z_j).
$P_-^{i,j}$	Probability of obtaining the negative pair (x_i, x_j) or (z_i, z_j).
$\mathcal{L}$	Loss function
$\mathbb{P}$	Parameter space. Includes θ, ψ
$\mathcal{P}$	A point in the Parameter space.
$\mathcal{P}_T$	Final Parameter state after T training iterations, $\mathcal{P}_T \in \mathbb{P}$
$\mathcal{P}_t$	Parameter state after t training iterations, $\mathcal{P}_t \in \mathbb{P}$
$\mathbb{G}$	Gradient space.
$\mathcal{G}$	A point in the Gradient space.
$\mathcal{G}_t$	Gradient state after t training steps, $\mathcal{G}_t \in \mathbb{G}$
$\mathcal{I}$	Mutual Information
η	Step Size / Learning Rate
$\mathcal{H}$	Hessian Matrix
$\mathcal{E}$	Joint Embedding Space
L	Lipschitz Constant
I	Input image
M	Segmentation mask corresponding to I
$\mathcal{S}$	Support set
$\mathcal{Q}$	Query set
$\mathbf{P}$	Prototypes
$\mathbf{p}$	pixel location
N_{pro}	Number of prototypes
$\mathcal{M}_c$	Cosine similarity matrix
$\mathcal{M}_{prob}$	Prototype probability matrix

Chapter 1

Introduction

In the era of machine learning, the rapid growth of novel techniques was boosted by the rapid increase in the amount of available digital data. Although initial techniques such as Naive Bayes, k-Nearest neighbour or Support Vector Machine were suitable for tasks comprising small datasets such as Iris (Unwin and Kleinman, 2021), MNIST (Deng, 2012), etc., the advent of more complex data forced researchers to opt for alternative techniques. While low-dimensional numerical feature-based data are still adequately learnt by these methods, the advent of high-dimensional data posed problems. Algorithms such as Naive Bayes or SVM are still used to this date, but they require careful modification to the basic framework to perform satisfactorily (Do, 2021; Do and Le Thi, 2022) on large datasets such as ImageNet (Deng et al., 2009). Furthermore, the scalability of these techniques also posed a problem with the applicability to more complicated data, hence the need for such modifications. More data meant that the time to train these models also increased, hence the need for faster computational alternatives which could be parallelized and sped up on modern computing hardware.

Another phenomenon that came forth due to high-dimensional data in methods like support vector machines is data piling, which resulted in the projections of the data in low-dimensional space being identical resulting in many samples being treated as support and therefore affecting generalization (Marron et al., 2007). Similar problems were also reported for the k-Nearest-Neighbour methods (Kouiroukidis and Evangelidis, 2011). The basic assumption in Naive Bayes that all features are independent of each other, helps to alleviate the curse of dimensionality. However, the assumption itself does not hold in high-dimensional data. Very often, the real variation in the data lies in a low-dimensional manifold in the high-dimensional space. Several dimension reduction techniques like PCA (Maćkiewicz and Ratajczak, 1993), tSNE (van der Maaten and Hinton, 2008) or UMAP (McInnes et al., 2018), it is possible to extract the low dimensional manifold features from the high-dimensional data. However, computation time and resources increase exponentially with increasing data. Dimensionality reduction techniques like PCA aim to maximize the variation in the data by linearly mapping high-dimensional data to a low-dimensional one. ISOMAP or Locally Linear Embedding (Saul and Roweis, 2001) aims to retain the local data properties intact and can identify the underlying non-linear structure in the data. LDA (Fisher, 1936) is similar to PCA but is capable of handling the curse

of dimensionality. tSNE (van der Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018) aim to preserve the mutual relationships between data points and learn non-linear mapping too. tSNE and UMAP are also immune to the curse of dimensionality. Nevertheless, we can say that the primary cause of the curse of dimensionality is the use of distance functions for high-dimensional data.

Unlike the above techniques, LeCun et al. (1989a,b) introduced the paradigm of learning parameters using backpropagation in the image domain. While the concept of biological neuron-inspired networks (McCulloch and Pitts, 1943; Fukushima, 1969) is not new, the adoption of neural networks gained traction much later. Researchers started adopting neural networks only at the beginning of the millennium (Bengio and Bengio, 2000). For high-dimensional images, convolutional neural networks were first proposed in Lecun et al. (1998) where the authors used gradient-based learning for handwritten digit classification. But they became popular only after the groundbreaking performance of AlexNet (Krizhevsky et al., 2012) on the ImageNet challenge. From thereon, there has been rapid development in the neural network architectures. Advancement in neural network architecture coupled with advancement in computational hardware boosted the growth of the deep learning paradigm.

The widespread application of deep learning to various tasks such as object detection, segmentation, and tracking led to more efforts in constructing larger curated datasets, as the data-hungry nature of the deep learning techniques became evident. An increase in the complexity of the data led to further innovations in deep learning architectures, resulting in the invention of Transformer architectures (Dosovitskiy et al., 2021), borrowing the philosophy of attention (Vaswani et al., 2017) from the domain of natural language processing.

All techniques discussed so far fall under the canopy of *Supervised* learning, as these techniques required ground-truth human annotated labels for learning the necessary features or representations or patterns. Contemporary to the works introducing the supervised learning techniques, researchers also investigated methods or frameworks which enabled learning of representations or patterns without ground truth human annotated labels. These methods fall under the canopy of *Unsupervised* learning. These included various methods like clustering and dimensionality reduction. *Clustering* requires multi-dimensional features which were either extracted using techniques such as SIFT (Lowe, 1999), HoG (Dalal and Triggs, 2005), SURF (Bay et al., 2006) before the application of neural networks. After neural networks were popularised, neural network encoders were also used for extracting features. Clustering too, like the supervised techniques, suffers from the curse of dimensionality, as the concept of distance becomes less precise in high dimensions. However, there have been attempts to deal with the issue using various modifications.

In current times, researchers use *Autoencoders* or pre-trained *Encoders* to efficiently map high-dimensional data into a low-dimensional manifold. These low-dimensional features can then be used for various tasks. Autoencoders also serve as a dimensionality reduction technique. In autoencoders (Gallinari et al., 1987; Bourlard and Kamp, 1988; Hinton and Zemel, 1993), the decoder is generally discarded and only the encoder is used for purposes such as dimensionality reduction, clustering, etc. However, it is expected that if we use the decoder with a random latent vector as input, it will be able to generate

a properly reconstructed picture. However, auto-encoders generally tend to overfit and are unable to generalize. To remedy this issue, variational autoencoders (VAE) (Kingma and Welling, 2014) were used, which are a probabilistic variation of the autoencoder. By learning to reconstruct the input data and matching the latent distribution to a prior, the VAE effectively learns a probabilistic representation of the data distribution. Like VAE, diffusion models (Ho et al., 2020) are also trained by minimizing the variational lower bound of the likelihood of the data distribution. However, diffusion models can be supervised or unsupervised.

The domain of *Deep learning* continues to evolve, and with the rapid growth in the computational hardware industry, this growth will be further enhanced. Deep learning techniques have found their way into the daily life of the common man through various applications, either through cell phones or personal computers. However, with the increase in the diversity of applications, the need for annotated data availability also grows. While it is possible to collect data which are uncurated and unannotated, the possibility of learning and inferring decisions based on those data is not plausible. Historically, in the last decade, under the canopy of unsupervised learning, researchers have developed another paradigm of learning algorithms, commonly termed as, *Self-supervised* learning. These learning algorithms allow the researchers to deal with the issues of overfitting and transferability of pre-trained supervised models to small-scale high-dimensional datasets.

In the subsequent sections of this chapter, we discuss the different facts of machine learning, like supervised and unsupervised, followed by the challenges associated with the self-supervised learning paradigm. This is followed by a discussion of the motivation and contribution of the thesis. Finally, we discuss the organization of the thesis.

1.1 Facets of Machine Learning

Based on principles, machine learning algorithms can be primarily categorized into two types: (a) *Supervised*, and (b) *Unsupervised* learning. Supervised learning algorithms utilise supervisory signals or information to learn the mapping between the data samples and the associated ground truth labels. These supervisory signals or information are generally obtained from the set of labels paired with the data samples. For a classification task, any supervised learning algorithm maximizes the log-likelihood of the conditional categorical data distribution $p(y|x)$, which models the probability of discrete class labels $y \in \{1, \ldots, K\}$ given inputs x, which is equivalent to minimizing the cross-entropy loss given the data $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{\aleph}, y_{\aleph})\}$, where $\aleph$ denotes the number of samples in the dataset. Here, the distribution $p(y|x)$ is parameterized by a model (e.g., a neural network with softmax output) that estimates class probabilities, and maximizing the log-likelihood corresponds to aligning predicted probabilities with the true class labels. Similarly, for regression tasks, the objective is to maximize the log-likelihood of the conditional data distribution $p(y|x)$, modelled as a continuous probability distribution (e.g., Gaussian). Maximizing the Gaussian log-likelihood parametrized by the mean and variance reduces the regression problem to minimizing the mean squared error loss.

In unsupervised learning, however, the ground-truth labels y are not available and is a type of machine learning where the algorithm learns patterns or structures from unlabeled

data without predefined output labels or targets. Common unsupervised learning tasks like clustering, dimensionality reduction or anomaly detection require learning the data distribution to effectively learn useful patterns from the data. Thus, the objective can be stated as maximizing the log-likelihood of the data distribution $p(x)$, given the data $\mathcal{D} = \{x_1, x_2, \ldots, x_\aleph\}$.

Reinforcement learning (RL) is another paradigm of machine learning, different from the above two. In reinforcement learning, an agent learns to make decisions through learning to maximize a cumulative numerical reward, instead of trying to learn any data distribution or any underlying structure from the data. One of the main challenges in reinforcement learning is the trade-off between exploration and exploitation which helps in learning optimal policies. RL has been used in several applications like playing games (like AlphaGo, Chess, DOTA2), robotics, autonomous driving, recommender systems, and most recently in training and fine-tuning large language models through RLHF (Reinforcement Learning from Human Feedback) (Christiano et al., 2017; Ziegler et al., 2019).

Another paradigm of learning, which has become pretty popular in the last decade is *Semi-supervised* learning. This method of learning combines both supervised and unsupervised learning and utilizes both labelled and unlabeled data. Semi-supervised learning techniques learn the structure of the data from the labelled data and further explore the manifold of representations through unsupervised techniques. Semi-supervised method is particularly useful when labelled data is scarce or expensive to obtain. In several real-world scenarios like web page data or social media content, there is abundant unlabelled data but limited labelled data. In such cases, semi-supervised learning can prove to be useful. There are several methods used for semi-supervised learning, such as pseudo-labelling, self-training or label propagation.

The latest addition to the list of paradigms is Self-supervised learning. It is a sub-category of unsupervised learning. Unlike, semi-supervised learning, where the pseudo-labels for the unlabeled data are obtained after training on the labelled data, for self-supervised learning the pseudo-labels are generated or defined by the researchers, and then utilized to learn representations. However, there are other methods which do not require assigning pseudo-labels to the data samples and learning representations from the unlabeled data like contrastive learning or joint embedding architectures. Thus, instead of labelled data, SSL uses pretext tasks (self-defined tasks) to extract supervision from raw data. The goal is to pre-train models on vast unlabeled data to capture universal patterns (e.g., language structure, visual features), which can later be fine-tuned on labelled data for specific tasks. This paradigm has found significant applications in natural language processing, computer vision-based tasks, speech recognition, multi-modal learning, etc. In short, self-supervised learning has revolutionized the domain of artificial intelligence by enabling modern foundational models like GPT (Achiam et al., 2023; Brown et al., 2020), DALL-E (Ramesh et al., 2021), and Diffusion models (Wei et al., 2023; Ho et al., 2020) to learn from massive unlabeled datasets (e.g., internet-scale text or images) and power generative AI and transfer learning.

1.2 Self-Supervised Learning

Self-supervised learning (SSL) is a machine learning paradigm in which the model learns representation from the data itself without requiring manually labelled data. It is a subset of unsupervised learning that utilises the inherent structure of the data to create supervisory signals. Self-supervised learning algorithms utilise the aforementioned supervisory signals obtained from the data to learn underlying semantic patterns of the data and are used or fine-tuned later for specific tasks like classification, segmentation, generation, etc. The key concepts in Self-Supervised Learning can be summarised as follows:

- **Pretext Tasks**: These are manually designed tasks where the model is trained to learn representations from the data. Common pretext tasks include
  - **Predicting missing parts**: Image inpainting (Pathak et al., 2016), predicting the relative position of two image patches (Doersch et al., 2015).
  - **Temporal ordering**: Predicting the order of frames in a video (Fernando et al., 2017) or the next word in a sentence, predicting missing clips (El-Nouby et al., 2019), etc.
  - **Data transformation recognition**: Recognizing the type of transformation applied to the data (e.g., rotations (Gidaris et al., 2018), colourization (Larsson et al., 2017), etc.).
  - **Reconstruction**: Reconstructing the input from the noisy or masked version of the same. In natural language processing (NLP), models like BERT (Devlin et al., 2019) are trained to predict masked words in a sentence, enabling them to learn context and meaning. Similar principles are also applied in masked autoencoder-based frameworks (He et al., 2022) as well.
  - **Generative Pre-training**: Models like GPT (Radford, 2018; Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023) generate text based on a given prompt, learning the structure and nuances of the language in the process. General adversarial networks (Goodfellow et al., 2020) and diffusion models (Ho et al., 2020) except class-conditioned ones can also be treated as self-supervised generative pre-trained models.
  - **Contrastive loss optimization**: Learning representations such that similar samples are mapped closer, while dissimilar samples are mapped farther apart. SimCLR (Chen et al., 2020a) and MoCo (He et al., 2020) are popular contrastive learning frameworks.
  - **Feature decorrelation**: Decorrelating the feature dimensions results in learning representations by discarding redundant information.
- **Learning Representations**: The main goal of SSL is to learn useful representations of data that can be transferred to downstream tasks. For instance, a model trained with SSL on images should learn features that can be useful for tasks like object detection or classification. The aforementioned goal is achieved through utilising supervisory signals obtained from the data and learning the underlying patterns in the data through the pretext tasks mentioned in the previous item.

- **Transfer Learning**: The representations learned via SSL are often fine-tuned on a small amount of labelled data for specific tasks. This reduces the dependence on large labelled datasets, which can be expensive and time-consuming to obtain. Common examples include transferring models pre-trained on the ImageNet dataset (Deng et al., 2009) using an SSL framework for tasks like object detection or segmentation. One of the complex examples includes fine-tuning a large language model like BERT (Devlin et al., 2019) pre-trained using masked language modelling (Devlin et al., 2019) for specific tasks like a chatbot or personal assistant.

Over the last few years, there have been significant applications of Self-Supervised Learning in various domains. Some notable applications are as follows:

- **Medical Image Analysis**: Most medical datasets are small compared to natural image datasets. Thus, supervised training or fine-tuning pre-trained networks on small datasets may lead to overfitting issues or the destruction of hierarchical co-adaptation between the network layers. To avoid such issues, self-supervised pre-training helps to adapt the pre-trained representations to the target data distribution. The availability of only a few publicly available medical image datasets of size comparable to ImageNet makes the problem of self-supervised learning more challenging on medical images. There have been applications of SSL techniques in the medical image analysis domain for different data like Chest X-rays, Abdominal MR and CT scans, skin lesion images, ultrasound, echocardiogram, endoscopy, retinal images, etc. Some noteworthy applications involving medical images that have gained traction include tasks like medical image registration, segmentation, classification, medical report generation, etc.
- **Natural Language Processing (NLP)**: Models like BERT use SSL by predicting masked words in a sentence. This pre-training step helps the model learn the structure and semantics of the language, which can be fine-tuned for various NLP tasks. Several modern large language models are also pre-trained using SSL techniques like masked language modelling.
- **Speech and Audio Processing**: SSL can be applied to audio data by predicting masked parts of the audio waveform or learning to distinguish between different augmentations of the same audio clip. Contrastive learning methods are also used for audio and speech applications.
- **Computer Vision**: SSL is used to train models on large datasets of unlabeled images. Tasks like image inpainting (predicting missing parts of an image) or predicting the rotation of an image helps the model learn useful visual features. Furthermore, maximizing the mutual information between two views of the same samples using a contrastive learning framework has also enhanced the capability of SSL models in representation learning.
- **Multi-modal applications**: Paired data like video or captioned images can be used for learning multi-modal representations. Recently, Vision Language models are also examples of multi-modal applications of SSL. Recently, medical report generation from medical images is also one notable application of such methods.

Now, let us discuss what are the benefits of Self-supervised Learning that make us choose it over supervised learning for representation learning and subsequent tasks.

- **Reduced Dependence on Labelled Data**: SSL leverages the abundant unlabeled data, reducing the need for extensive and costly labelled datasets. Utilising the unlabeled data, SSL frameworks are capable of extracting useful representations which can be used to initialise networks in the downstream task. Furthermore, pretext training on small datasets also enhances the adaptability of the transferred weights to the downstream tasks.
- **Improved Generalization**: As the SSL models are trained without labels, these models often generalize better to new tasks because they learn richer and more versatile representations. This results due to incorporation of versatile augmentations during pretext training, thereby increasing the variance in the data, consequently boosting generalization.
- **Preventing over-fitting in data-scarce scenarios**: Generally in medical imaging analysis, ImageNet pre-trained weights are used. However, the distribution of ImageNet and the target medical dataset are generally different. This requires fine-tuning and may destroy the hierarchical co-adaptation in the features. However, if the network is pre-trained on the target medical dataset, the distribution becomes similar to that of the target task. This will prevent over-fitting in the downstream task in data-scarce scenarios by enhancing the adaptability of the transferred weights to the downstream task.

Relationship between SSL and Human Psychology The use of self-supervised learning algorithms in Orhan et al. (2020) links this paradigm to the field of psychology. Using ego-centric videos, the authors aimed to determine the origin of the ability of infants to perform basic discriminatory tasks. As defined in Cattell (1963), fluid general ability refers to the ability to adapt to new situations, whereas crystallized general ability refers to those cognitive abilities in which skilled judgment has become crystallized. In Horn (1965), we learn that fluid intelligence and crystallized intelligence in fact are co-dependent. In light of the theory of intelligence proposed in Cattell (1963) and later extended in Horn and Cattell (1966) and Horn (1965), we can formulate self-supervised learning as the fluid ability to learn novel knowledge bases or representations based on self-designed cues or in other words, without cues from external agents. The transfer of self-supervised pre-trained weights to other downstream tasks resembles the utilization of fluid intelligence to help in building up crystallized intelligence, as described in Horn (1965). The problem statement in Bridle et al. (1991) of learning to classify samples without prior knowledge or labelled examples, fits the analogy of infants having the ability to discriminate between animal classes, which in turn, points to the similarity of this learning paradigm with fluid intelligence or ability in the field of psychology.

Despite several benefits of the self-supervised learning paradigm, there are several challenges which occur during the implementation of self-supervised learning frameworks. In the following subsection, we discuss a few challenges associated with this learning paradigm.

1.2.1 Challenges of Self-Supervised Learning in Medical Image Analysis

In this subsection, we discuss the challenges associated with self-supervised learning that we have dealt with in this thesis. There are several challenges in self-supervised learning which can be treated as general challenges, such as the dependence of the quality of representations on the pretext tasks, sensitivity to data augmentation, need for large computational resources, need for a large amount of unlabeled data, learning of low-rank representations causing dimensional or complete collapse, overfitting in pretext tasks, etc.

In addition to the aforementioned general challenges associated with SSL frameworks, we discuss the challenges that we have taken care of in this thesis. Those are discussed as follows:

- **Limited data**: As medical data is hard to collect and annotate by expert personnel, we often find limited publicly available medical data. This makes training any model from scratch on small medical image datasets difficult. From the discussion of works on medical imaging modalities, we can observe that many of those pieces of work are on MR and CT images, but applications of SSL to other medical imaging modalities are limited. Even though it is possible to pre-train a network on a medical image dataset using a self-supervised framework, it will not be transferable to all medical imaging modalities due to the limitation in the diversity of the medical image data.
- **Context-based pretext tasks do not learn context-invariant representations**: In context-based pretext tasks, such as rotation prediction or jigsaw puzzle solving, the representations learnt are not context-invariant or transformation-invariant (Misra and van der Maaten, 2019). When the representations learnt are not invariant to the transformation being applied, the network learns to solve the predictive pretext task using features dependent on the context. This phenomenon results in representations that are not invariant with the context being modelled, resulting in the learning of redundant representations or shortcut solutions. As a consequence of the aforementioned phenomenon, the quality of representations degrades.
- **Limited improvement in Context-based tasks**: Another intriguing question in self-supervised learning is, whether self-supervised pre-training improves performance by adapting the learnt representations to the target dataset in data-scarce scenarios. As the data distribution of the ImageNet (Deng et al., 2009) and the target datasets differ, the need for fine-tuning in the downstream task arises. If the scale of the target dataset is small, then the issue of overfitting and destruction of co-adaptation between hierarchical features also occurs. However, there are only a few studies which explores the effect of self-supervised pretraining on medical image datasets.
- **Dependence on Uniformity-Alignment tradeoff**: In self-supervised contrastive learning, the negative pairs play an important role in influencing the quality of representations and consequently the downstream performance. However, in conventional contrastive learning (van den Oord et al., 2018; Chen et al., 2020a; He et al., 2020), only mutual information between the positive pair samples is maximized. While attempts have been made to modify the effects of hard negative samples to improve

performance, none of the works attempts to explicitly incorporate the mutual information of the negative pairs in the equation. This eliminates an important element in controlling the uniformity-alignment trade-off and the downstream performance.

- **Limited understanding of convergence in contrastive learning**: Furthermore, in self-supervised learning, the phenomenon of convergence is also not well understood. While there has been ample research for studying the convergence of gradient descent algorithms in supervised learning scenarios, the same cannot be said for self-supervised learning.
- **Limited understanding of the role of temperature in contrastive learning**: In addition to that, the temperature hyper-parameter plays an important role in controlling the attraction and repulsion force between samples in positive and negative pairs, respectively. This phenomenon, in turn, controls the optimization process as a whole. However, the role and effect of the temperature hyper-parameter have not been widely explored. This limits the proper exploration of the role of temperature and also limits a possible improvement in the quality of representations.
- **Pretext and downstream task disparity**: The most significant part of the self-supervised framework is the pretext task. All the above-mentioned challenges are dependent on the pretext task used for pre-training the SSL model. The pretext task and the downstream task often differ in their objective. While the network architecture remains mostly unchanged, the optimal representations for the pretext and downstream tasks differ with the change in objective. Furthermore, as the objective differs, the representations learnt in the pretext task and the representations which are optimal for the downstream task also differ. Hence, there is an alignment issue between pretext and downstream representations arising due to pretext and downstream task disparity.

In addition to the above-mentioned challenges, there are some other challenges associated with application of SSL in medical image analysis, such as domain-specific generalizability, limited freedom of data augmentation, noise and artefacts in medical images harming supervisory signals, lack of benchmarking, etc.

Despite the challenges, self-supervised learning has been greatly successful in various domains such as computer vision, natural language processing, speech and audio processing, etc. Self-supervised learning has become the go-to tool for researchers for obtaining representations from unlabelled data, as it provides a better initialization which is better adapted to the downstream task data than any pre-trained networks trained on a completely different data distribution. Hence, outperforming supervised models with a little fine-tuning serves as a better trade-off in that scenario. Furthermore, due to the ability of SSL frameworks to capture different levels of generalized hierarchical information, representations learnt through SSL pre-training are also easily transferable to other tasks and data as well. The recent technical advances like large language models, and vision language models owe their rapid growth to SSL. In vision tasks, contrastive learning models, denoising-based generative models like diffusion models and reconstruction-based models also have succeeded in outperforming supervised baseline models satisfactorily. In this

regard, we have explored the topic of self-supervised learning and its application in medical image analysis in this thesis. The detailed motivation of the thesis is discussed in the following section.

1.3 Motivation of the Thesis

In this section, we discuss the motivations that drive the works done in this thesis. The motivations are discussed as follows:

- **Learning of Context-invariant representations**: One of the challenges of context-based pretext tasks, as already discussed in Misra and van der Maaten (2019), is that the representations learnt were not context-invariant and contain redundancy. If the representations contain redundancy then they are effectively low-rank, consequently harming the representation learning capacity of the model. Hence, this challenge motivates us to develop SSL frameworks which can learn context-invariant representations by decoupling spatial representations and improving representation learning.
- **Handling data-scarce scenarios**: Conventionally, researchers use ImageNet pre-trained weights obtained from supervised classification tasks as initialization for any task. Representations learnt using context-based tasks can serve as good initialization for the downstream tasks, but lag behind contemporary supervised algorithms in downstream performance. Furthermore, in data-scarce scenarios, effective fine-tuning is also difficult as the data distributions of the source and target datasets differ. This may destroy the hierarchical co-adaptation in the pre-trained weights. Hence, this motivates us to study the effect of self-supervised pre-training on top of ImageNet pre-trained weights which helps to maximize the utilization of the weights learnt in the pretext task and enhance downstream task performance in data-scarce scenarios.
- **Binary Contrastive Learning**: Contemporary InfoNCE-based contrastive SSL frameworks deal with two types of pairs, positive and negative. The basic principle of contrastive learning translates to pulling the similar samples closer while pushing the dissimilar samples farther apart. Hence, our motivation is to design a novel binary contrastive learning framework which does the same by simply predicting a pair as either positive or negative.
- **Mutual Information optimization**: Contemporary contrastive SSL frameworks maximize the mutual information between the positive pairs, whereas the mutual information between the negative pairs is not explicitly incorporated into the formulation. However, the negative pairs play an important role in controlling the uniformity-alignment trade-off which subsequently determines the quality of representations. This motivates towards the modification to the binary contrastive framework which balances the effect of positive and negative pairs. This modification not only maximizes the mutual information between the two samples in a positive pair but also explicitly minimizes the mutual information between the two samples in a negative pair.

- **Adaptive temperature for better representation learning**: The temperature hyper-parameter plays a significant role in controlling the attraction and repulsion between the different types of pairs in contrastive learning, and consequently, controls the alignment-uniformity trade-off. The temperature hyper-parameter also controls the hardness of the false negative pairs and influences the representation learning process. However, work on this aspect of contrastive learning is sparsely explored. This motivates the development of a temperature scaling function which modulates the temperature hyper-parameter adaptively for each pair and enhances the representation learning.

- **Minimizing pretext and downstream task disparity**: The pretext task and the downstream task often differ in their objectives. The pretext task also controls the quality and type of representations that will be learnt and transferred to the downstream task. It plays a significant role in determining the adaptability of the transferred features. To fully utilise the representations learnt in the pretext task, it should be aligned with the nature of representations required in the downstream task. However, little study has been done to minimize the pretext and downstream task disparity. This gap in the research landscape motivates our attempt to minimize the pretext and downstream task disparity by developing a self-supervised framework with the pretext task being the same as the downstream task.

1.4 Contribution of the Thesis

The contributions of the thesis are as follows:

- **Design of Context-invariant SSL**: We devise a semi-parallel convolutional architecture to decouple the context between each patch of the jigsaw, and prevent learning of low-level signals like discontinuities, edges and blank spaces. Furthermore, the features from each of the semi-parallel convolutional branches are concatenated along the feature dimension for better generalization. We further propose novel convolutional blocks to enhance pretext representation learning. We also study the effect of increasing the number of parameters results in better downstream performance.

- **Novel Binary Contrastive Framework**: We propose a novel contrastive learning framework based on the noise contrastive estimation principle for binary classification scenarios. This aforementioned framework is based on the simple strategy of classifying the pairs as positive or negative. We also investigate whether self-supervised pre-training on top of ImageNet pre-trained weights boosts performance in the downstream tasks.

- **Mutual information optimization through the lens of uniformity and alignment**: Taking the binary contrastive learning framework as the baseline, we further improve it to mitigate the effects of the positive and negative pair imbalance by using the variational perspective of the maximum likelihood formulation of the binary contrastive framework. We find that the proposed modification of the binary contrastive framework not only maximizes the mutual information between positive

pair samples but also explicitly minimizes the mutual information between the negative samples. This leads to a better uniformity-alignment trade-off and subsequently influences the convergence of SSL frameworks.

- **Convergence analysis of SSL frameworks**: The convergence phenomenon in SSL frameworks is not well investigated in the current research landscape. We analyse the convergence of SSL frameworks by locally satisfying the Polyak-Lojasiewicz inequality on the non-convex loss landscape.
- **Estimation of Optimal temperature hyper-parameter**: The temperature hyper-parameter plays an important role in controlling the uniformity-alignment trade-off and subsequently the quality of representations. We take a basic assumption that the gradient of the contrastive loss with respect to the cosine similarity of a negative pair should be positive, as with the decrease in the contrastive loss, the cosine similarity of a negative pair should also decrease in an ideal case. Based on this assumption, we solve a first-order ordinary differential equation obtained from expanding the gradient term. The resulting equation estimates the ideal temperature scaling function which controls the temperature hyper-parameter value as a function of the cosine similarity of any pair.
- **Novel Self-supervised One-shot Segmentation**: To minimize the disparity between the representations for the pretext and downstream task, we propose a self-supervised one-shot segmentation framework in the pretext task for a one-shot segmentation downstream task. This objective helps the network to learn representations which are aligned to the downstream task and achieve satisfactory performance without fine-tuning. We propose a correlation-weighted prototype aggregation step which incorporated both the local and global context information to efficiently segment the desired regions in abdominal medical scans.

1.5 Layout of the Thesis

This thesis focuses on proposing novel self-supervised frameworks for learning representations from data without utilising human-annotated labels. This thesis consists of eight chapters and we discuss the organization of the thesis as follows. Different medical imaging modalities like magnetic resonance (MR), computed tomography (CT), X-ray, skin lesion images, etc. are used for our empirical validation in this thesis.

- In **Chapter** 1, initially we discuss the current scenario of machine learning and its many facets. Then we proceed to discuss about self-supervised learning and its challenges in medical image analysis. Finally, we also discuss the motivation, contribution and layout of the thesis in this chapter.
- **Chapter** 2 deals with literature survey. In this chapter, we discuss the different types of self-supervised frameworks categorised based on the type of principles. We have also critically analysed the different foundational self-supervised frameworks. Furthermore, we review the works that use SSL frameworks for learning representations from medical image data along with categorisation based on the imaging

modality. In addition to that, this chapter provides a discussion of the advantages and shortcomings of those pieces of work.

- **Chapter** 3 proposes novel context-based self-supervised learning frameworks that utilize a jigsaw puzzle-solving strategy to learn representations from medical visual data. The jigsaw puzzle-solving strategy is used to learn better relations between the different spatial features of an image, in addition to the proposed frameworks aiding in learning context-invariant representations. The proposed frameworks also prevent learning of low-level signals as well. For empirical justification of the proposed framework, we used MRNet (Bien et al., 2018) and KneeMR (Štajduhar et al., 2017) datasets and achieved performance at par with the contemporary supervised baselines.

- In **Chapter** 4, we propose a novel binary contrastive learning framework based on the basic principle of classifying pairs in a contrastive learning scenario as positive or negative. Additionally, this chapter investigates whether self-supervised pre-training on top of ImageNet pre-trained weights boosts performance in the downstream tasks or not. For empirical evidence, we conducted the experiments on the MRNet dataset and achieved results which outperformed the contemporary supervised baseline.

- In **Chapter** 5, we improve the binary contrastive learning framework by deriving the learning objective from the variational perspective of the maximum likelihood formulation of the binary contrastive framework. We also investigate the convergence of the SSL framework by locally satisfying Polyak-Lojasiewicz inequality on the non-convex loss landscape. Through our experiments on both small-scale (CIFAR) and large-scale (ImageNet) benchmark datasets, we show that the proposed framework outperforms the contemporary self-supervised learning frameworks. We furthermore showed that the proposed frameworks also outperform the state-of-the-art SSL frameworks on transfer learning tasks on several medical image datasets like MURA (Rajpurkar et al., 2017), Chaoyang (Zhu et al., 2022), ISIC Skin Lesion dataset (ISDIS, 2016), etc.

- To improve uniformity-alignment trade-off and subsequent representation learning, in **Chapter** 6, we propose a temperature function for controlling the temperature hyper-parameter in contrastive learning. For this purpose, we solve a first-order differential equation obtained from the gradient of the InfoNCE loss with respect to the cosine similarity of a negative pair. Through empirical evidence, we show that the proposed framework can improve the baseline SimCLR framework such that it outperforms the contemporary self-supervised learning frameworks on benchmark small-scale and large-scale datasets. The proposed framework was also evaluated on transfer learning tasks on medical image datasets like MURA (Rajpurkar et al., 2017), Chaoyang (Zhu et al., 2022), ISIC Skin Lesion dataset (ISDIS, 2016), etc. From the experimental results, it can be observed that the proposed framework outperforms the contemporary state-of-the-art SSL frameworks on several of those datasets.

- **Chapter** 7 proposes a novel correlation-weighted prototype aggregation step to efficiently incorporate local and global contextual information for self-supervised one-shot segmentation in the pretext task. This effectively minimizes the pretext

and downstream task disparity and helps in learning representation in the pretext task, which is better aligned to the downstream task. The empirical justification of the proposed framework was obtained through experiments on the abdominal multi-organ MR dataset (CHAOS (Kavur et al., 2021)), and the proposed framework performed at par with the contemporary self-supervised few-shot segmentation frameworks without fine-tuning.

- **Chapter** 8 concludes the major findings and significant contributions of the work with possible directions for future research.

In the next chapter, we will discuss and analyze the relevant pieces of literature in the self-supervised learning domain.

Chapter 2

Literature Survey

In this chapter, we present a brief overview of the research landscape of self-supervised learning and various applications on different modalities of medical images.

2.1 Introduction

Deep learning has facilitated substantial progress in various applied fields such as signal processing, computer vision (CV), natural language processing (NLP), and others. Scaling the architecture has been a go-to strategy for researchers to improve performance along with algorithmic novelty. The application of deep learning in medical image analysis is also not an alien thought and instances of such on brain MRI, knee MRI, colonoscopy videos, chest X-ray images, mammograms, etc. are plentiful.

One of the drawbacks of supervised deep learning methods is the requirement for large amounts of labelled data, without which supervised deep learning models tend to overfit and fail to generalize. Supervised deep learning models require long training periods to achieve satisfactory performance even when a large amount of data is available. To deal with the overfitting problems, researchers employ transfer learning techniques to train supervised deep learning models on small-scale datasets. Deep learning models trained on large-scale datasets like ImageNet (Deng et al., 2009) or MS-COCO (Lin et al., 2014) are often used as pre-trained models in many applications, even if there is a domain mismatch between the pre-training and target dataset. Labelled data is limited or hard to obtain in real-life applications such as medical image analysis. Furthermore, annotations from domain experts are time-consuming and labour-intensive. Machine learning paradigms like semi-supervised and self-supervised learning have emerged to deal with such issues in supervised learning. In this study, we will focus primarily on self-supervised learning algorithms. In later sections, we will also discuss the applications of self-supervised learning strategies in different medical image modalities for representation learning.

In the following subsections, at first, we will discuss different studies on Self-supervised algorithms and frameworks in Sec. 2.2. The section starts with a discussion on the classical

approaches, such as context-based pretext tasks (Sec. 2.2.1), followed by the clustering-based frameworks (Sec. 2.2.2 and paired embedding-based methods (Sec. 2.2.3), which includes both contrastive and non-contrastive methods. This section ends with a discussion on works which combine techniques from more than one type of pretext task and hence cannot be categorized into any single one. Next, in Sec. 2.3 we present a discussion on different studies of SSL in the medical image analysis domain. These studies are subdivided with respect to the imaging modality, such as MRI & CT (Sec. 2.3.1), Ultrasound (Sec. 2.3.2), Endoscopy (Sec. 2.3.3), Radiographs (Sec. 2.3.4), Retinal images (Sec. 2.3.5), histopathology (Sec. 2.3.6), echocardiogram (Sec. 2.3.7) and skin images (Sec. 2.3.8). Finally, we end this survey with a summary and conclusion in Sec. 2.4.

2.2 Studies on Self-supervised Algorithms and Frameworks

In this section, we categorize and analyse the research on self-supervised learning into three primary categories: context-based, clustering-based, and paired embedding-based pretext tasks. Each category is further subdivided based on the underlying principles employed by the works. Within these subsections, we focus on discussing the foundational and notable contributions that have shaped each principle-based subcategory.

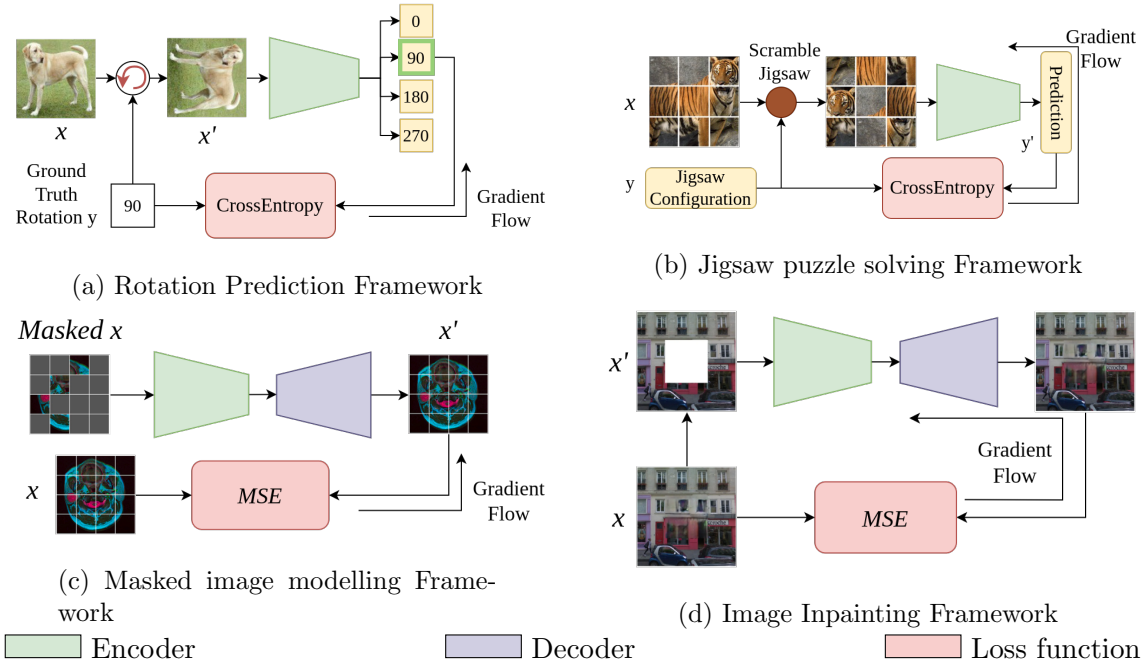
2.2.1 Context Based Pretext Tasks

In this subsection, we primarily discuss the different context-based pretext tasks used in self-supervised pre-training. The context-based pretext tasks can be categorized primarily into seven types, namely (Spatial) Context Encoding, geometrical transformation prediction, Jigsaw Puzzle Solving, Colorization, Spatio-Temporal Context Prediction (used primarily for videos), Masked Image Modeling, and Masked Video Modeling. We will start our discussion with geometrical transformation prediction-based pretext tasks and then continue with the others. The abstract illustration of a few foundational context-based frameworks is depicted in Fig. 2.1.

2.2.1.1 Context Encoding

The strategy of context encoding for unsupervised feature learning was first introduced in Doersch et al. (2014), wherein the authors use the prediction of the position of a single patch as a supervisory context prediction task to learn object clusters for unsupervised object discovery. In a later work, Doersch et al. (2015), the authors employed AlexNet (Krizhevsky et al., 2012) to classify the position of a patch with one reference patch sampled apriori as context from the same image.

In Pathak et al. (2016), image inpainting has been used as a generative context encoding task. A DCGAN (Radford et al., 2016) based generative pipeline, with a joint loss, consisting of l_2 and adversarial loss, and conditioned on the masked regions only was used and showed considerable improvement over the nearest-neighbor based image inpainting method.



(a) Rotation Prediction Framework

(b) Jigsaw puzzle solving Framework

(c) Masked image modelling Framework

(d) Image Inpainting Framework

FIGURE 2.1: Illustration of Context-based Frameworks. "Gradient flow" indicates the direction along which the parameter gradient propagation occurs, that is, starting from the loss and through the network.

2.2.1.2 Geometrical Transformation Prediction

In Agrawal et al. (2015), the authors used the prediction of the transformation of the camera from pair images as a pretext task. Another work along similar lines was presented in Jayaraman and Grauman (2017), where the objective is to learn the ego-motion equivariance from image pairs selected from ego-motion videos. Jayaraman and Grauman (2017) also combined contrastive loss to enforce equivariance between image pairs with supervised classification loss.

In the work RotNet (Gidaris et al., 2018), the authors trained the network to predict the rotation of the images, forcing the network to learn the semantic features in the objects to effectively classify the orientation of the dominant features in the objects. To improve representation learning, Feng et al. (2019) combined rotation prediction with instance discrimination task to learn both equivariant and rotation invariant representations.

Rotation prediction was also applied to videos in Jing and Tian (2018), where the authors adopted a 3DCNN architecture (3DRotNet) to account for both spatial and temporal information in videos. In another work, by using a 3D Convolutional Auto-Encoder to predict future frames, in addition to classifying rotation applied on the frames, Kumar et al. (2021) outperforms the vanilla 3DRotNet by a considerable margin on video retrieval benchmark tasks.

2.2.1.3 Jigsaw Puzzle Solving

The primary objective of jigsaw puzzle-solving pretext tasks is to learn spatially invariant contextual information by learning to predict the order of arrangement of the patches. The conventional approach involves dividing the image into several square patches and numbering them in raster order. After jumbling the position of the patches, the network is trained to predict the order of arrangement or generate the arranged image.

To the best of our knowledge, Noroozi and Favaro (2016) was the first to use jigsaw puzzle solving as a pretext task. After dividing each input image into 9 patches, the authors used 9 separate encoders for extracting representations from the patches and also prevented learning of low-level artefacts. The combined output from the 9 encoders was then merged to predict the jigsaw arrangement.

In Kim et al. (2018), where the authors combined jigsaw puzzle solving, image colourization, and image inpainting in a multi-task learning problem for self-supervised learning of representations from images. In Wei et al. (2019), the authors take an innovative approach of iteratively reorganizing the patches by probabilistic assignment of patches to a particular position, as well as, optimizing the relativistic position assignment of any two patches, until convergence. In another innovative approach Chen et al. (2021b), the authors clustered the patches and used the prediction of cluster assignment for each patch as the pretext task. JigsawGAN (Li et al., 2022) combined the flow information between the input and the rearranged patches with a generative reconstruction task for representation learning.

VideoJigsaw (Ahsan et al., 2019) applied the jigsaw puzzle solving strategy in Noroozi and Favaro (2016) to videos. After dividing each frame into 2×2 grid and rearranging all the patches over all the timesteps, a network is trained using a curriculum learning strategy to predict the arrangement of the patches by learning both spatial context and temporal order of the events. In Kim et al. (2019), a 3D CNN classifies the 3D jigsaw with a separate 3D CNN encoder for each space-time cube to prevent learning of low-level cues.

2.2.1.4 Colorization

To convert a grayscale image into a coloured one, the network needs to learn the semantic features of each class of objects. This fundamental principle is utilized to devise pretext tasks for self-supervised representation learning.

In the first, Iizuka et al. (2016) used a combination of self-supervised pre-training with a colourization task and supervised classification for representation learning. Using a different approach, Zhang et al. (2016) treats the colourization problem as a classification problem, wherein the *ab* colour space is quantized into 313 classes. Next, the work Larsson et al. (2017) is heavily inspired by Larsson et al. (2016), where the framework predicts a colour histogram at each location of the pixels.

2.2.1.5 Video Spatiotemporal Contextual Prediction

The objective of video representation learning is to learn both spatial and temporal features. We will discuss the works done in this sub-domain and also categorize them according to the pretext strategy used.

SSCAP (Wang et al., 2022c) uses image-based context-based SSL representation learning frameworks for feature extraction from the frames of a video for subsequent co-occurrence action parsing for action segmentation. In Luo et al. (2020), a number of separate clips from a video were generated in order and one clip was randomly removed. The network is trained to predict the option which has been used to alter the removed clip, in light of the context of the other clips. In Duan et al. (2022), a ranking-based framework was used to learn semantic and temporal information from unlabeled videos.

In one of the first papers on self-supervised video representation learning Wang and Gupta (2015), the objective is to map the features of two different instances of the same object obtained by tracking improved density trajectory or SURF feature points over multiple frames of a video. Building upon the work done in Doersch et al. (2015) and Wang and Gupta (2015), Wang et al. (2017) utilized transitive relation invariance, constituting both inter-instance relations between different object instances of similar appearance and intra-instance relations between identical objects at different timesteps for visual representation learning.

In Misra et al. (2016), the representations are learned by classifying if a tuple consisting of 3 frames sampled from a high-motion window in the video, is in the correct order or not. Similarly, sequential variation in visual features has been used to learn representations in OOO (Fernando et al., 2017), Lee et al. (2017), Skip-Clip (El-Nouby et al., 2019) as well. In OOO (Fernando et al., 2017), the network is tasked with predicting the index of a clip with the incorrect order of frames among several clips from the same video. Whereas in Lee et al. (2017), the network is tasked to predict the order in which 4 frames in the input are arranged. In Skip-Clip (El-Nouby et al., 2019), the objective of ranking clips based on a given context clip as plausible future clips of the given context is used as self-supervision.

2.2.1.6 Masked Image Modeling

Over the last few years, Masked Image Modeling (MIM) has developed as a new and popular avenue of research under the canopy of self-supervised learning. The principle is the same as in contextual information learning like Doersch et al. (2015) and Pathak et al. (2016). The concept is inspired from masked language modeling frameworks like BERT (Devlin et al., 2019) in the domain of Natural Language Processing (NLP). Compared to the previous context-based pretext tasks which mostly focus on the local contextual information, MIM-based frameworks utilise both local and global information which leads to better representation learning for dense prediction-based downstream tasks.

Initial works like Chen et al. (2020), and BEiT (Bao et al., 2021) introduced the concept of MIM in SSL. In BEiT (Bao et al., 2021), a ViT-based encoder is used to predict the visual tokens of the masked image patches, where the tokens are obtained from a discrete

variational auto-encoder (dVAE) (Rolfe, 2017). iBOT (Zhou et al., 2022a) uses a self-distillation based MIM framework to train a target encoder with a momentum-updated encoder as the online tokenizer. SimMIM (Xie et al., 2022) uses raw pixel value regression as the objective function when reconstructing the original image from the masked input image. SimMIM also uses a learnable mask token vector to replace each masked patch, like BERT.

Masked Auto Encoders (MAE) (He et al., 2022) are another prime example of instance-based SSL frameworks which utilise masked image modelling. Unlike BERT (Devlin et al., 2019), MAE uses an encoder-decoder architecture. MAEs use a high masking ratio of 70-80% for optimal performance. The encoder only processes a small portion of the patches, whereas the decoder processes both the input latent representations and the mask tokens to learn generalized representations.

CAE (Chen et al., 2023c) combines masked representation regression and masked patch reconstruction. Another such work BootMAE (Dong et al., 2022) uses the output of a momentum-updated target encoder as a target for feature prediction with a multi-scale pixel regression objective. DMAE (Wu et al., 2023) aims to learn robust representations by reconstructing the original input from noise-corrupted images. In CIM (Fang et al., 2023), to generate the corrupted image, a dVAE (Rolfe, 2017) and a small pre-trained BEiT (Bao et al., 2021) are used as the frozen tokenizer and generator, respectively. GAN-MAE (Fei et al., 2023), uses MAE as the corrupt image generator and a GAN-like discriminator is used to classify the output image from MAE as fake or real. In DiffMAE (Wei et al., 2023), the denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) is combined with MAE (He et al., 2022) for visual representation learning.

Contrastive MAE (CMAE) (Huang et al., 2023b) uses both masked patch/frame reconstruction to learn locally sensitive semantic representations and contrastive loss optimization to maximize the similarity between different representations and also to learn the discriminative relation between different images. Similarly to CMAE, RePre (Wang et al., 2022a) also uses a contrastive objective to maximize the representational similarity between different augmented videos. However, it uses a specialized reconstruction decoder for each level of the multiple hierarchy features obtained from the ViT encoder.

ConvMAE (Gao et al., 2022a) introduces a hybrid convolution-transformer encoder architecture. The convolution layers are primarily used for the high-resolution embeddings, while the transformer layers are used for the low-resolution embeddings. ConvNeXtv2 (Woo et al., 2023) uses a sparse convolutional encoder and a convolutional block decoder as a fully-convolutional MAE, inspired by the ConvNeXt (Liu et al., 2022c) architecture.

2.2.1.7 Masked Video Modeling

Masked video modelling (MVM) extends the concept of masked image modelling (MIM) to videos. Several recent works on masked video modelling are discussed below.

BEVT (Wang et al., 2022b) uses a VideoSwin (Liu et al., 2022d) transformer as a shared image and video encoder, but a separate decoder for image and video. MaskFeat (Wei et al., 2022) uses masked feature prediction to learn visual features for video understanding,

but uses a dVAE codebook such as BEiT for tokenization. VideoMAE (Tong et al., 2022) employs tube masking to handle temporal redundancy and correlation, to learn representations from videos. VideoMAEv2 (Wang et al., 2023b) further scales VideoMAE by using a dual masking strategy to make video understanding more efficient. In addition to an encoder mask, VideoMAEv2 also uses a decoder mask following MAR (Qing et al., 2023). Feichtenhofer et al. (2022) extends MAE (He et al., 2022) to video understanding, by dividing a video into a regular grid of patches in space-time and using a high masking ratio to reduce the computational complexity in the encoder.

2.2.2 Clustering-based Frameworks

One of the first pre-training approaches using K-Means algorithms to learn a patchwise feature dictionary was presented in Coates and Ng (2012). DeepCluster (Caron et al., 2018) utilises k-means clustering algorithm to generate pseudo-labels from features extracted by the convolutional neural networks, which are then used for the cross-entropy loss-based classification task for representation learning. DeeperCluster (Caron et al., 2019) adds context-based pretext tasks with DeepCluster for better pre-training. In ODC (Zhan et al., 2020) the pseudo-labels evolve alongside the parameters preventing rapid change to the pseudo-labels, with two separate memory banks for samples and centroids, eliminating the need for an extra feature extraction step.

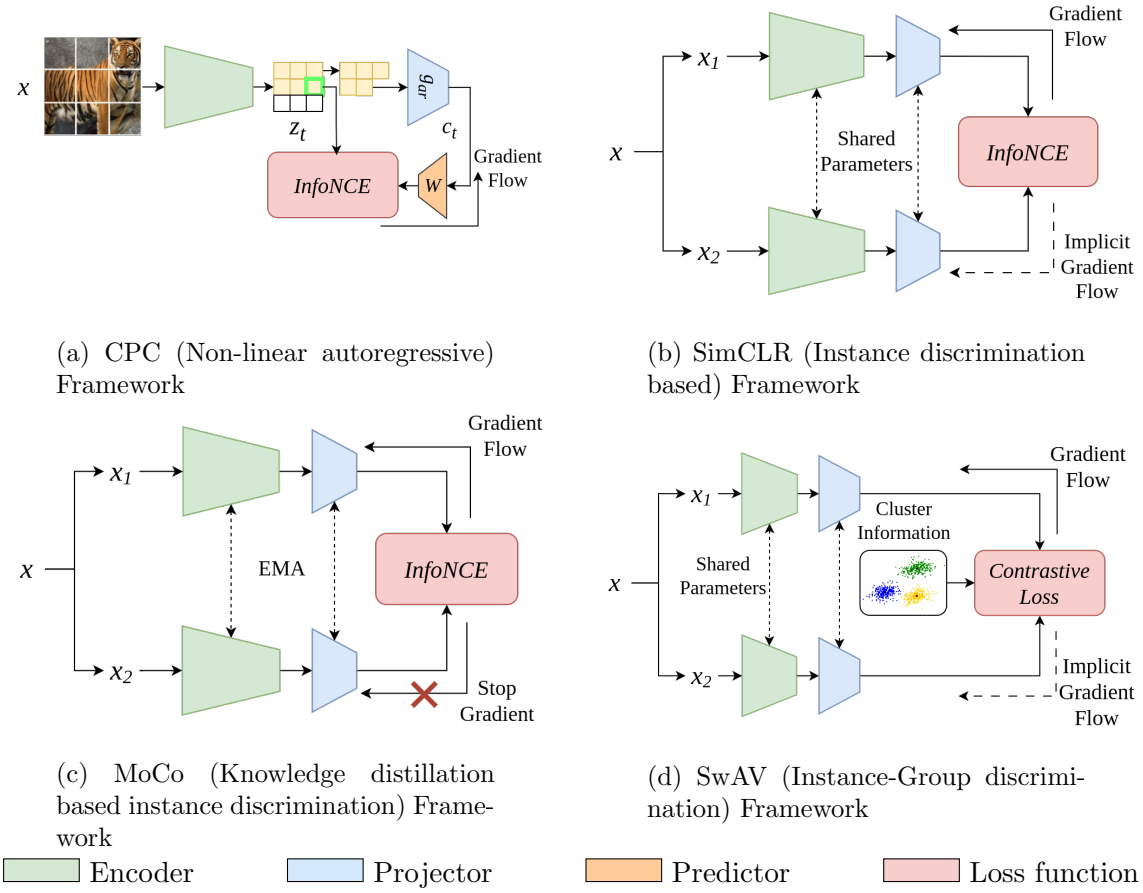
JULE (Yang et al., 2016) proposed a recurrent network-based unsupervised framework that combines agglomerative clustering for pseudo-label generation and subsequent classification. CoKe (Qian et al., 2021) utilises an online constrained K-means algorithm to compute pseudo-labels and cluster centres to capture the global distribution of the data.

DEC (Xie et al., 2016) uses an autoencoder for parameter initialization and a KL-divergence based clustering and parameter optimization step. IDEC (Guo et al., 2017) incorporates the auto-encoder in the DEC framework itself to preserve the local embedding structure. SDMVC (Xu et al., 2023) further scales IDEC to multiple views by using an autoencoder for each view and clustering all views to facilitate global discriminative feature learning.

RIM (Krause et al., 2010) improves Bridle et al. (1991) by using a regularizer to ensure proper clustering. IMSAT (Hu et al., 2017) uses RIM (Krause et al., 2010) for the clustering step followed by an information maximization step for representation learning.

2.2.3 Paired Embedding Based Pretext Tasks

While the context-based pretext tasks laid the foundation of SSL in the early years of its development, those frameworks failed to compete with the performance of models using transfer learning mechanisms with supervised pre-trained weights. However, a new family of frameworks using information from paired embeddings to optimize a specific objective or loss function, and learning optimal parameters in the process, provided a huge leap in performance. We can categorize these frameworks primarily into two categories: (1) Contrastive and (2) Non-contrastive. We will discuss several foundational and current state-of-the-art frameworks in both categories below.



(a) CPC (Non-linear autoregressive) Framework

(b) SimCLR (Instance discrimination based) Framework

(c) MoCo (Knowledge distillation based instance discrimination) Framework

(d) SwAV (Instance-Group discrimination) Framework

FIGURE 2.2: Illustration of Contrastive Frameworks. "Implicit gradient flow" means the gradient flow is not restricted for the second views (x_2) of the sample x in frameworks like SimCLR and SwAV. In MoCo, the second view x_2 is passed through the momentum updated encoder, hence, no gradient flows through the same, which is represented by "Stop gradient". "EMA" denotes Exponential moving average.

2.2.3.1 Contrastive Learning Frameworks

In the literal sense, contrastive learning can be considered as learning by contrasting different samples. The primary objective of contrastive learning frameworks is to discriminate between dissimilar samples, and closely map similar samples. While supervised triplet loss-based contrastive learning frameworks have been around for a long time (Chopra et al., 2005; Weinberger and Saul, 2009), the use of contrastive loss in an unsupervised setting was first observed in works like Hyvärinen and Morioka (2016); Sermanet et al. (2017). In Fig. 2.2, we illustrate the abstract representation of some foundational contrastive frameworks which serve as the baseline for the current advances in self-supervised learning. In the following sections, the contrastive learning frameworks are categorized primarily into two types, non-linear autoregression-based and instance discrimination-based.

(i) Non-linear Autoregressive Frameworks

The paradigm of contrastive self-supervised learning (SSCL) frameworks received a huge

boost with the advent of CPC (van den Oord et al., 2018) as it introduced the principle of InfoNCE, primarily based on the principle of noise contrastive estimation (NCE) (Gutmann and Hyvärinen, 2012). CPC uses a patch-based autoregressive style predictive framework. The work showed that optimizing InfoNCE is synonymous with maximizing the mutual information between the input and its corresponding representation. The principle of CPC (van den Oord et al., 2018) was later used again in CPCv2 (Hénaff et al., 2020). CPCv2 improved CPC by increasing model capacity, applying layer normalization, using more context information for predicting patch embeddings, and patch-based augmentations. CPC was further improved in RPC (Tsai et al., 2021), where the authors improved the training stability, sensitivity to minibatch, and downstream performance by eliminating the logarithm of contrastive loss and using an additional l_2-regularization.

(ii) Instance Discrimination Frameworks
Instance discrimination frameworks contrast instances or samples with each other for learning representations. There has been a huge development in this direction, along with several derivatives of this primary framework, like instance-instance discrimination, knowledge distillation-based instance discrimination, instance-group discrimination, multimodal contrastive learning, etc.
(a) Instance-Instance Discrimination: An innovative perspective on instance-instance discrimination was presented much before CPC in the work Dosovitskiy et al. (2014), where the pretext task is simply a multiclass classification task, which involves learning to classify the transformed set of images.

Concurrently with CPC (van den Oord et al., 2018), Inst. disc. (Wu et al., 2018) proposed a novel instance discrimination framework based on the principle of NCE. Wu et al. (2018) treated each instance as a separate class of its own and maintained a memory bank to construct a non-parametric softmax classifier for self-supervised representation learning. This work laid the foundation for several SSCL frameworks for vision tasks. Concurrent to Wu et al. (2018), AMDIM Bachman et al. (2019) presented a self-supervised version of Deep InfoMax Hjelm et al. (2019) and also expanded the architecture to incorporate multiscale features allowing more efficient mutual information maximization between samples in positive pairs.

The basic framework of LA (Zhuang et al., 2019) is primarily based on a clustering step to identify close and background neighbours from a momentum-updated memory bank and then apply a contrastive loss-based local aggregation metric. PIRL (Misra and van der Maaten, 2019) uses a formulation of contrastive loss following Hadsell et al. (2006). In fact, the formulation of Wu et al. (2018) is a special case of PIRL. PIRL was the first to point out that the representations learnt in the contemporary context-based frameworks are not transformation-invariant leading to learning of redundant representations.

SimCLR (Chen et al., 2020a) is the first SSL framework to not use a memory bank for the formation of negative pairs. Instead, it used a large batch size and showed that increasing the number of negative pairs improves performance. It also emphasized the role of augmentation and non-linear projectors in the quality of representations in SSL. SimCLRv2 (Chen et al., 2020b) further found that increasing the number of parameters results in better representation learning for semi-supervised and fine-tuning performance, if the labels are fewer. The basic framework of SimCLR has also been used in contrastive

video representation learning frameworks such as CoCLR (Han et al., 2020), TCLR (Dave et al., 2021), CVRL (Qian et al., 2021), etc.

NNCLR (Dwibedi et al., 2021) incorporates multiple positive instances in SimCLR (Chen et al., 2020a) by sampling positive samples from the manifold neighbourhood of each instance represented by a support set queue as in MoCov1 (He et al., 2020) or MoCov2 (Chen et al., 2020c). SNCLR (Ge et al., 2023) employs an additional weight calculation step to compute the correlation between the support instance and the sampled neighbours which are then used in the N-pair contrastive loss.

DCL (Yeh et al., 2022), eliminated the negative-positive-coupling effect in InfoNCE loss which proved harmful to the learning efficiency in contrastive frameworks, resulting in a significant improvement in performance without the requirement of large batch size, such as in SimCLR (Chen et al., 2020a) or momentum encoding in MoCo (He et al., 2020).

SSL-HSIC (Li et al., 2021c) proves that InfoNCE is an approximation of the proposed framework with a variance-based regularization and proposes an HSIC (Hilbert-Schmidt Independence Criterion) bottleneck inspired loss, computed using an estimator provided by Gretton et al. (2005). TiCo (Zhu et al., 2022) introduced a novel contrastive framework by using a squared contrastive loss instead of the widely used InfoNCE loss. It also encourages the representations of negative samples to be orthogonal. A concurrent work by HaoChen et al. (2021) explores spectral contrastive learning, where the authors use the population augmentation graph to effectively partition the same into sub-graphs, which are representative of fine-grained sub-classes of the actual classes in the downstream task. In reality, this work uses a learnable spectral decomposition component of the embeddings to learn the most important eigenvectors to maximise linear probe performance.

In literature, there was a dearth in metrics to measure the quality of representations. Wang and Isola (2020) analyzed self-supervised representation learning using two metrics, alignment and uniformity, and studied the relationship of the two metrics with the downstream performance. An intriguingly similar work was also presented in Ye et al. (2019). Moon et al. (2022) empirically discovered that the alignment and uniformity are directly correlated with the downstream performance of both instance-level and dense-level downstream tasks.

DenseCL (Wang et al., 2020) proposes a pixel-wise contrastive learning framework using a convex combination of the pixel and image level InfoNCE loss for dense representation learning, specifically for downstream tasks like semantic or instance segmentation, depth estimation, etc. SetSim (Wang et al., 2021) tries to improve DenseCL by finding the correspondence set of features for the query feature vectors from the queue, as well as, pixel-level features from the feature attention maps. Recent works like Shen et al. (2023) use asymmetric masking to generate positive samples and stop-gradient to prevent the collapse of representations with InfoNCE as the loss function.

CMC (Tian et al., 2020) proposed one of the first attempts to extend self-supervised contrastive learning to more than two views using a self-supervised version of N-pair loss objective (Sohn, 2016). In addition to the modified objective, CMC also uses a memory bank similar to the above frameworks. Another major contribution of CMC is the introduction of the ImageNet100 dataset in the SSL domain, which is a subset of the 100

class of the original ImageNet1K dataset (Deng et al., 2009). Similar to CMC, MIL-NCE (Miech et al., 2020) uses multiple positive pairs for multi-modal representation learning. In a recent work Hu et al. (2024), the authors use a combination of pseudo-label guided positive pair sampling step, feature and cluster correlation maximization, and divergence minimization-based clustering for multi-view self-supervised learning.

One of the oldest foundational works on SSL is Bridle et al. (1991). In this work, the output is used as the probability distribution over the class label, a discrete random variable. The objective is to maximize the difference between the entropy of the average of the outputs (referred to as fairness) and the average of the entropy of the outputs (referred to as firmness). Maximizing the entropy of the average of the outputs prevents the dimensional collapse of representations, while minimizing the average of the entropy of the outputs prevents collapse of the representation to a single point in the latent space. This principle is the backbone of all self-supervised learning frameworks.

ReLIC (Mitrovic et al., 2021) explores SSL from a causal perspective with content and style as latent variables. This work argues that the conditional distribution of the class representations given the content should remain invariant under style changes and uses the KL divergence as a regularizer along with contrastive loss.

(b) Knowledge Distillation-based Instance Discrimination Frameworks: Another significant work using a memory bank for negative sample mining was presented in MoCo (He et al., 2020). MoCo uses a momentum-updated encoder to extract representations to store in the fixed-size memory bank. All the representations in the memory bank acts as negative samples, while a positive pair is obtained by pairing two differently augmented versions of a sample. This extends the idea of knowledge distillation but uses the distribution of the likelihoods of an augmented version of the query to supervise the retrieval of another version of the query from a pool of representations.

MoCov2 (Chen et al., 2020c) attempted to further improve MoCo (He et al., 2020) by incorporating two design properties of SimCLR, that is, non-linear projection head, and stronger data augmentation. Another attempt to further scale up MoCo (He et al., 2020) was presented in MoCov3 (Chen et al., 2021c) which ditched the memory bank as it showed minimal gain when used with a large batch size. In MoCov3, an additional prediction head was also used in the online encoder, following BYOL (Grill et al., 2020). Zhu et al. (2021) proposes a simple yet effective feature transformation, which creates both "hard positives" and "diversified negatives" to enhance the training with MoCov2 as the baseline framework. VideoMoCo (Pan et al., 2021) has presented an extension of the MoCo framework to videos, where the authors used temporal adversarial learning to augment videos.

In a concurrent work, CAST (Selvaraju et al., 2021) uses Deep-USPS (Nguyen et al., 2019) to identify salient regions and a constrained cropping augmentation method to avoid the inclusion of noisy background regions. Using MoCo (He et al., 2020) as the baseline, CAST uses an additional attention loss to base predictions on correct regions, as contrastive methods often use wrong regions to match query and key images.

CLSA (Wang and Qi, 2021) uses distribution divergence minimization using representations from a memory bank, in addition to contrastive loss for better representation learning with MoCov2 as the baseline. CO2 (Wei et al., 2021) also uses the MoCov2 framework as

a baseline along with a KL divergence-based consistency regularization loss in addition to contrastive loss.

(c) Mitigating Sampling Bias: One noteworthy characteristic of contrastive SSL is that samples with the same label were paired into negative pairs. Debiased CL (Chuang et al., 2020) attempted to remove the negative sampling bias from the available positive sample pairs only. Robinson et al. (2021) proposed to utilise hard negative samples to improve the generalization of self-supervised learning frameworks, and improve downstream performance.

(d) Instance-Group Discrimination: Clustering-based pretext tasks aim to adapt clustering algorithms for the generation of pseudo-labels for end-to-end training of visual features.

In G-SimCLR (Chakraborty et al., 2020), the authors use an autoencoder to extract representations for a k-means clustering-based pseudo-label generation from each batch before applying contrastive loss. IIC (Ji et al., 2019) simply utilises mutual information to learn representations from data in an unsupervised manner, discarding instance-specific details and also preventing the assignment of all samples to a single cluster using individual cluster assignment entropy maximization.

PCL (Li et al., 2021a) formulated the proposed framework as an Expectation-Maximization algorithm. After clustering the features from the momentum encoder, negative sample prototypes are sampled, and finally, the NCE loss is optimized. ProPos (Huang et al., 2021) optimizes the MSE loss between a sample and its Gaussian distributed positive neighbours, in addition to the objective proposed in PCL. In another recent work, MUGS (Zhou et al., 2022b) uses three levels of granularity for representation learning, instance level, local group level, and global group level.

Simply stated, clustering requires assigning samples to each of the clusters. When a uniformity condition is applied to the assignment problem, it can be treated as an optimal transport problem. In SeLa (Asano et al., 2020b), the first step is the same as the previous clustering-based pretext task, that is, cluster-based pseudo-label assignment and cross-entropy-based optimization. The second step involves *Sinkhorn-Knopp* algorithm-based transport polytope computation (Cuturi, 2013). SwAV (Caron et al., 2020) uses clustering to compute prototypes which are used to predict codes of one view from another using *Sinkhorn-Knopp* algorithm. SMoG (Pang et al., 2022) further improves SwAV by adding a group-level discrimination branch to it. MIRA (Lee et al., 2022) also improves SwAV by not using the equipartition constraint, rather it constrains the marginal entropy by mutual information regularization. SEER (Goyal et al., 2022) explores the challenges of scaling the pre-training architectures using SwAV as the baseline framework and also addresses some of the engineering challenges and complexity of training at this scale.

Instead of using K-means for generating pseudo-labels like in DeepCluster (Caron et al., 2018), ODC (Zhan et al., 2020) or CoKe (Qian et al., 2021), SCAN (Van Gansbeke et al., 2020) first uses a pretext task to learn representations and then obtains the clusters using neighbour sampling with the prior knowledge of the number of classes in the dataset. The parameters are then fine-tuned again using cross-entropy loss similar to SeLa (Asano et al., 2020b). Similar to SeLa and SwAV, MDRA (Cheng et al., 2023) also uses optimal

transport for relationship alignment, which is another term used to assign samples to prototypes. However, each feature vector is decomposed into subgroups along the feature dimension, and the procedure of relationship alignment is applied to each subgroup.

One fundamental shortcoming of SeLa, SwAV, and SMoG is the assumption of uniform distribution over the prototypes which ensures the possibility of converging to degenerate cases but ignores that real-life datasets are skewed. SeLaVi (Asano et al., 2020a) presents a solution to this using a permutation matrix in the energy equation of the Sinkhorn-Knopp algorithm (Cuturi, 2013), which sorts the prototype entropies to account for imbalanced data distribution.

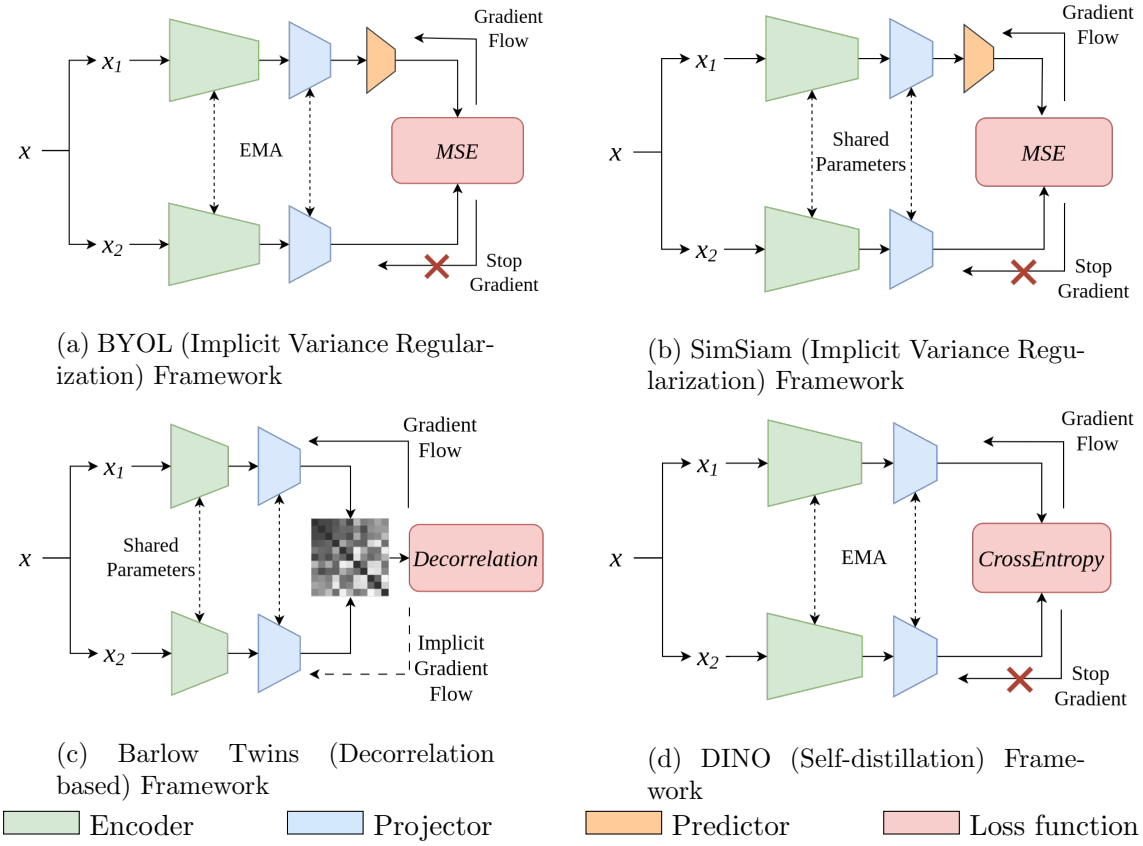
(e) Multi-modal Contrastive Learning: AVTS (Korbar et al., 2018) uses triplet contrastive loss instead of InfoNCE loss, with distance-based self-supervised synchronization between video and audio modalities. CBT (Sun et al., 2019) uses the principle of masked language modelling on videos and paired textual information separately, as well as cross-modal contrastive learning to maximize the mutual information between visual and textual modes of information.

AVID (Morgado et al., 2021) uses cross-modal contrastive learning and within-modal positive discrimination using a sampled positive and negative set. Different from AVID, STiCA (Patrick et al., 2021) uses both cross-modal and within-modal contrastive loss for multi-modal representation learning. ACC Ma et al. (2021) uses cross-modal contrastive learning with MoCo as the baseline, but the output from the momentum-updated encoder of the other modality is used as the target representation. MVCGC (Huo et al., 2021) uses positives sampled from the pool of samples of the other modality as hard positives and also incorporates cross-modal information by optimizing N-pair contrastive loss.

VATT (Akbari et al., 2021) uses MIL-NCE (Miech et al., 2020) for representation learning from video, and text, and an additional NCE loss for video and audio. FNACL (Sun et al., 2023) uses false-negative suppression and true-negative enhancement in the contrastive learning framework for sound source localization tasks.

2.2.3.2 Non-Contrastive Frameworks

The frameworks which do not explicitly use the contrastive loss for self-supervised pre-training are categorized under Non-contrastive learning. In principle, these frameworks discard the negative pairs and only use the positive pairs in the self-supervised pre-training phase. One of the first non-contrastive frameworks can be traced back to De Sa (1993) and de Sa (2014), where the primary objective is to minimize the disagreement between the information from two different modalities. An innovative yet simple approach of using uniformly sampled noise from the l_2 unit sphere as fixed target representations to avoid collapse in self-supervised learning was presented in NAT (Bojanowski and Joulin, 2017). In Fig. 2.3, we present the abstract illustration of some foundational non-contrastive frameworks for better understanding. In the following sections, the non-contrastive frameworks are categorized primarily into the following five types, implicit variance regularization-based, knowledge distillation-based, decorrelation-based, spectral decomposition-based, and distribution divergence minimization-based.



(a) BYOL (Implicit Variance Regularization) Framework

(b) SimSiam (Implicit Variance Regularization) Framework

(c) Barlow Twins (Decorrelation based) Framework

(d) DINO (Self-distillation) Framework

FIGURE 2.3: Illustration of Non-contrastive Frameworks. "Implicit gradient flow" means the gradient flow is not restricted for the second view (x_2) of the sample x in frameworks like SimCLR and SwAV. In MoCo, the second view x_2 is passed through the momentum updated encoder, hence, no gradient flows through the same, which is represented by "Stop gradient". "EMA" denotes Exponential moving average.

(i) Implicit Variance Regularization-based Framework

The requirement of a large number of negative samples for instance discrimination-based contrastive learning led to the advent of negative-free contrastive learning methods. However, to prevent collapse or trivial solutions, it was necessary to have access to the statistics of the negative samples. Hence, instead of peeking into the batch dimension, researchers utilised the information available along each embedding dimension, which allowed these new frameworks to avoid collapse without explicit instance contrast. Regulating the variance along the embedding dimensions, allows the information to be distributed over the embedding dimensions, preventing dimensional collapse. The previous statement is supported by findings in Tian et al. (2021), where it is stated that the predictor used in works like BYOL (Grill et al., 2020) or SimSiam (Chen and He, 2020), behave as a whitening transform preventing dimensional collapse. Furthermore, Halvagal et al. (2023) also emphasises the role of predictor and stop-gradient in preventing collapse and through eigenspace analysis of the predictor, shows that both BYOL and SimSiam perform implicit variance regularization for asymmetric Euclidean and cosine losses.

(ii) Knowledge Distillation-based Frameworks

One of the foundational works in the negative-free contrastive learning literature is BYOL (Grill et al., 2020) which uses a student (online) - teacher (target) network architecture as in knowledge distillation, but the teacher (target) network learns from the past iterations of the student (online) network. The objective is to maximize the similarity between the representations predicted by the online network and the representations of the target network. The predictor MLP is essential to prevent the collapse of representations in BYOL. Initially, it was hypothesized that the collapse of representations was prevented because of the batch normalization (BN) layers used in BYOL and that the BN induced an implicit contrastive effect on the embedding representations. However, these hypotheses were rejected in Richemond et al. (2020). Alternatively, Tian et al. (2021) states that the representational dynamics are decoupled for Euclidean losses as in BYOL, and converge to finite eigenvalues in the predictor's eigenspace. As the eigenvalues correspond to the variance of the representations, the underlying mechanism constitutes an implicit form of variance regularization. ASCNet (Huang et al., 2021) uses BYOL as the baseline framework for video representation learning using both appearance and speed consistency as the objective. FlowE (Xiong et al., 2021) also uses BYOL for predicting the representations of another frame from one frame after applying flow transformation. In another work, MYOW (Azabou et al., 2021) combines a distance loss between a sample x and mined samples from the latent neighbourhood of the sample x, with the objective used in BYOL (Grill et al., 2020). MSF (Koohpayegani et al., 2021) added a positive sampling step from a large memory bank to BYOL to improve consistency regularization. CMSF (Tejankar et al., 2021; Navaneet et al., 2022) improves MSF by utilizing different sources of knowledge like multi-modal embeddings to constrain the nearest neighbour search space.

SimSiam (Chen and He, 2020) uses an architecture similar to the online encoder in BYOL but without the momentum updated encoder. Instead, SimSiam adopts an alternating optimization problem by using a stop gradient to prevent collapse. Halvagal et al. (2023) show that SimSiam performs an implicit variance regularization when using the stop-gradient, without which the eigenvalues in the predictor's eigenspace diverge or collapse of representations occurs. BraVe (Recasens et al., 2021) also uses SimSiam as the baseline framework but for learning multimodal representation from videos. DenseSiam (Zhang et al., 2022b) adds a dense pixel-wise similarity loss and a region-based contrastive loss to SimSiam for dense representation learning.

Self-distillation is similar to knowledge distillation (Hinton et al., 2015) in supervised learning, but without *a priori* teacher network. In DINO (Caron et al., 2021) the parameters of the teacher network are generally obtained from the momentum encoding of the parameters of the student network over training iterations. The student network is trained to learn local-to-global correspondences by matching the probability distribution of both networks. EsViT (Li et al., 2022) observes that ViTs can automatically discover semantic correspondence between local regions, but the use of multi-stage ViT causes a loss of property. EsViT proposes a novel non-contrastive region matching pretext task to capture the local region dependency in the features. ReSSL (Zheng et al., 2021) presents a novel framework based on relational consistency between instances instead of explicit repelling-pulling of instances in contrastive learning by computing the similarity distribution of each instance with the representations in the memory bank and minimizing the KL-divergence to enforce the relation consistency between the two augmented views of an instance. A similar approach to ReSSL is also applied in ISD (Tejankar et al., 2020), however, instead

of a memory queue, it uses a collection of random samples to approximate the neighbourhood of the sample. Yun et al. (2022) improves DINO by mining positive patches from the neighbouring patches and using the aggregated representation as the target. In more recent work, DINOv2 (Oquab et al., 2023) scales self-supervised pre-training in terms of data and model size. It combines DINO (Caron et al., 2021) and iBOT (Zhou et al., 2022a) with the centering of SwAV (Caron et al., 2020) and KoLeo regularization (Sablayrolles et al., 2019). MST (Li et al., 2021c) also adopts a similar approach as DINO (Caron et al., 2021) and also optimizes a reconstruction loss from the student network by using an attention-guided masking strategy to mask out low response patches.

MSN (Assran et al., 2022) combines masked image modeling with the self-distillation framework. However, unlike DINO (Caron et al., 2021), MSN uses a set of prototypes to calculate the softmax probabilities. PMSN (Assran et al., 2023) relaxes the condition of uniform clustering in MSN by minimizing KL divergence with a power law distribution instead of a negative entropy term in MSN. CrOC (Stegmüller et al., 2023) uses DINO as the baseline framework and combines a clustering-based strategy by adding a representation centroid-based self-distillation pipeline with it.

(iii) Decorrelation-based Frameworks

Barlow Twins (BT) (Zbontar et al., 2021) presents an innovative approach without using any similarity-based loss. The framework proposed in BT uses an objective which can be understood as a form of information bottleneck, maximizing the variability of the representations over the dimensions, thereby preventing dimensional collapse, and also discarding redundant information arising from applied distortions or augmentations. The advantage of BT over InfoNCE-based frameworks is that it does not require a large batch size and benefits from large-dimensional embeddings. He and Ozay (2022) argue that Barlow Twins output whitened features, and explore the relation between collapse of representations and whitening of features, and the exponent of eigenspectrum which follows the power law decides the gap. Hua et al. (2021) explores the concept of collapse in BT and WMSE as baseline frameworks and claims to discover dimensional collapse in SSL as well. This work explores the role of feature decorrelation and proposes Shuffled Decorrelated BN for improved representation learning and prevention of dimensional collapse in SSL.

VICReg (Bardes et al., 2022a) also uses the decorrelation principle like BT but uses invariance maximization and variance regularization as the primary objectives. VICRegL (Bardes et al., 2022b) improves VICReg by adding location-based and feature-based matching of embeddings across both views of positive samples. SMT (Chen et al., 2023) is the simplest form of VICReg but uses a more restrictive linearity criterion for similarities.

(iv) Spectral Decomposition-based Feature Whitening

W-MSE (Ermolov et al., 2021) does not use a separate predictor like BYOL and also avoids collapse while using the same loss as BYOL, by using a Cholesky decomposition step to whiten the features along the batch dimension. ZeroCL (Zhang et al., 2022a) proposes a novel approach to self-supervised representation learning using independent invariance minimization along both batch and feature dimensions after instance-wise and feature-wise ZCA whitening, respectively. However, like W-MSE, ZeroCL involves a whitening step using Cholesky decomposition that is computationally expensive and is of the order

of $\mathcal{O}(n^3)$. ARB (Zhang et al., 2022b) proposes another new approach using orthonormal bases of one view of a sample as a target for the feature representations of the other view.

(v) Distribution Divergence Minimization

TWIST (Wang et al., 2021b) presents an interesting approach by minimizing the divergence between the probability distributions of two augmented samples, along with the entropy of each sample, along with a diversity term to ensure that representations of different samples are different to prevent collapse. This approach is similar to De Sa (1993); de Sa (2014). PMO (Luo and Wang, 2022) uses a matching operator to match input representations to a prior distribution, without the need for contrasting positive and negative samples, thereby preventing collapse.

2.2.4 Miscellaneous

There are other pieces of work which cannot be categorized in any of the above categories. Some of such notable works have been discussed below.

The design of the object counting problem as a pretext task in SSL can be seen in Noroozi et al. (2017), where the authors use a contrastive loss, instead of a regression loss, to prevent trivial solutions, a common problem in SSL.

Kolesnikov et al. (2019) studied the effect of CNN architectures with different SSL frameworks and found that: a) architecture choices may significantly affect performance in the self-supervised setting, b) the quality of learned representations in CNN architectures with skip-connections does not degrade towards the end of the model, c) increasing the size of the representation significantly and consistently increases the quality of the learned visual representations, and d) linear probing performance is sensitive to learning rate.

Gwilliam and Shrivastava (2022) proposed several metrics for benchmarking and analyzing self-supervised frameworks in their work. Basaj et al. (2021) proposes several visual probing tasks previously used in NLP to evaluate SSL frameworks. Ryali et al. (2021) addresses the problem of learning spurious correlations in SSL by investigating a class of simple, yet highly effective background augmentations, which encourage models to focus on semantically-relevant content by discouraging them from focusing on image backgrounds.

Islam et al. (2021) show that combining supervised loss with self-supervised contrastive loss improves transfer learning performance. Zhang et al. (2021c) also used contrastive loss as a regularizer along with cross-entropy loss in the downstream fine-tuning stage. Around the same time, Cole et al. (2021) investigates the impact of data quality and quantity, task granularity, and pre-training domain on the quality of representations learned in contrastive learning.

UnMix (Shen et al., 2022) employs image mixing methods like CutMix (Yun et al., 2019) and MixUp (Zhang et al., 2018) to implement an unsupervised counterpart of label smoothing in supervised learning to improve representation learning. MixCo (Kim et al., 2020) and i-Mix (Lee et al., 2021) are other concurrent works exploring the same image-mixing strategy for contrastive learning algorithms.

2.3 Studies of SSL in Medical Image Analysis

Self-supervised learning has demonstrated significant prowess in the domain of representation learning from medical imaging modalities. Unlike natural image datasets, medical image datasets are not abundant and are also expensive to annotate. Thus, self-supervised learning aids in adapting the parameters to the data distribution of the medical imaging modalities better than using ImageNet-pretrained weights. We have discussed studies supporting the above statement later in this section. Self-supervised learning frameworks have been applied to different medical imaging modalities for a host of tasks like classification, segmentation, anomaly detection, image reconstruction, etc.

Classification tasks include tasks like tumour classification from magnetic resonance or computed tomography images, injury classification from knee magnetic resonance images, identifying normal or abnormal tissues in histopathological images, etc. Similarly, SSL pre-training has also found application for segmentation tasks like tumour segmentation, tissue substructure segmentation, skin lesion segmentation, organ segmentation from whole-body magnetic resonance or computed tomography images, etc.

SSL is also used for reconstructing medical images from corrupted images or for denoising or removing artefacts in the same. Self-supervised learning is also applied for the super-resolution of medical images. Discovering patterns in medical data for anomaly detection by learning normal patterns without the need for extensive labelled data is also possible with self-supervised representation learning. Generation of synthetic data to deal with data scarcity is also made possible with the help of self-supervised pre-training. Furthermore, self-supervised learning can also be used for visualization of medical image datasets, multi-modal analysis of medical data, volumetric analysis of medical data such as fetal pose estimation, reference plane detection in ultrasound data, etc.

In the following subsections, we discuss several studies which utilise self-supervised frameworks for learning representation from different medical imaging modalities, such as MRI, CT, Ultrasound, echocardiogram, x-ray or radiographs, endoscopic images, retinal images, and skin images, for the variety of downstream tasks mentioned above.

2.3.1 MRI & CT

In this section, we discuss several works where the authors have used different SSL frameworks for representation learning from MRI and CT data. Here, we classify the works based on the different frameworks like context-based, instance discrimination-based, etc.

(a) Context-based Frameworks

Models Genesis (Zhou et al., 2021b) uses an image restoration-based task to learn image representations. Semantic Genesis (Haghighi et al., 2020) adds another image reconstruction-based pre-training stage to Models Genesis before the image restoration pipeline. Jana et al. (2021) also uses image restoration as a pretext task to learn representations from CT

images for liver fibrosis diagnosis. CaiD (Taher et al., 2022) reconstructs the original image from the corrupted version of the same image to learn context-aware representations in addition to an instance discrimination task.

Chen et al. (2019) uses restoration of corrupted images as a pretext task using a reconstruction-based framework. Sli2Vol (Yeung et al., 2021) uses a slice reconstruction-based strategy to learn representations to segment regions in 3D CT or MRI volume. Lu et al. (2020) and TractSeg (Lu et al., 2021) use pseudo-labels obtained from tractography for reconstruction to learn representations from fMRI data for segmentation. Demirel et al. (2021) uses a reconstruction-based framework proposed by Yaman et al. (2020) for simultaneous multi-slice image reconstruction task itself. SSL-LNE (Ouyang et al., 2021) also uses a reconstruction-based framework to learn the disease progression trajectory of individuals. Akçakaya et al. (2022) gives an overview of the different unsupervised methods used for biomedical image reconstruction. Sun et al. (2021) utilizes simulated artefacts obtained from the downsampling of MR scans to incorporate cortical thickness as anatomical guidance. In the testing phase, an iterative training stage is used to learn a site-specific segmentation network. Dong et al. (2021) reconstructs a fixed number of slices preceding and following the input slice of CT scans to learn representations. This work also uses a BYOL (Grill et al., 2020) as an auxiliary task in addition to the reconstruction-based task. Similarly to Dong et al. (2021), Alice (Jiang et al., 2023) uses a combination of masked image modelling and maximization of similarity between semantically aligned crops obtained using SAM (Yan et al., 2020) for representation learning. Chen et al. (2022b) show how masked image modelling outperforms traditional contrastive learning by speeding up convergence and greatly improving downstream task performance. TransMorph (Chen et al., 2022a) also uses a reconstruction-based approach for unsupervised image registration by predicting the deformation between fixed and moving images.

Another work due to Zhang et al. (2023b) uses a UNETR or Swin-UNETR-based 3D reconstruction-based framework for representation learning. SSPT-bpMRI (Yuan et al., 2023) uses a 3D UNet for reconstruction of 3D volume from an augmented sub-volume. The representations are then used for the detection and diagnosis of csPCa (prostate cancer). OneSeg (Wu et al., 2022c) learns the semantic correspondence between two different 2D slices from 3D CT scans using a reconstruction-based framework. Huang et al. (2022) uses symmetric positional encoding for Brain MR slices and 3D VHOG as targets for reconstruction from masked 3D voxels. VectorPose (Zhang et al., 2023b) uses boundary and voxel reconstruction, as well as spatial vector prediction, to learn spatial and anatomy-sensitive representations of 3D volumes. Mazher et al. (2024) uses a style transfer-based approach to learn representations from private datasets without compromising the data privacy of clients in a federated learning setting. Other works like Zhao et al. (2023) and M^3AE (Liu et al., 2023) also use a reconstruction-based framework.

Several works like Jog et al. (2016), Zhao et al. (2018), Xu et al. (2021), Zhao et al. (2021) use super-resolution as a pretext task. This allows the networks to learn contextual representations specific to the data, and also deal with the scarcity of high-resolution medical data.

PrimeGeoSeg (Tadokoro et al., 2023b) and Tadokoro et al. (2023a) synthesize 3D volumetric data using geometric shapes to emulate 3D MR scans and train a segmentation network

using the same pretext task. In another work, Zhang et al. (2023a) uses a synthetic tumour data generation pipeline for learning to segment brain tumours.

Spitzer et al. (2018) uses a Siamese architecture for representation learning from differently cropped versions of the input by maximizing the similarity between the two encoded inputs and also predicting the transformations applied on both inputs. Yang et al. (2020) uses the rotation and elastic prediction task as the source of self-supervisory signals in their framework, and the downstream segmentation module is also jointly trained with the self-supervision module. In addition to that, for disentanglement of appearance and content codes, the proposed framework also uses a modality-transfer generative module to learn cross-modal content-aware representations in an adversarial training way, and a self-reconstruction module.

Zhuang et al. (2019) propose an interesting approach by using Rubik's cube solving as a pretext task for representation learning from both MR and CT volumes. Zhu et al. (2020) further improves it by adding more augmentations for better representation learning. The application of jigsaw puzzle solving to medical image data was first done in Manna et al. (2022), where the authors used a semi-parallel architecture to predict the arrangement of the patches of Knee MR slices. This work showed the robustness of the jigsaw puzzle-solving strategy to data imbalance. However, the authors took a slice-based approach to learn representations from MRI scans. Each slice of the MR scan was divided into 9 patches, similar to Noroozi and Favaro (2016). For each patch, the authors used separate convolutional branches followed by a channel-wise aggregation step which enabled the network to decouple the representations from each patch, thereby preventing learning of low-level and redundant representations. The outputs were later merged and passed through custom convolutional blocks, to finally predict the arrangement of the patches. This work was further improved in SKID (Manna et al., 2023), where the architecture was modified to learn better representations. To deal with the 3D nature of MR scans, the authors used a ConvLSTM-based (Shi et al., 2015) classifier in the downstream task, and kept the encoder parameters frozen.

Taleb et al. (2021) uses a multimodal jigsaw puzzle-solving task, where each patch is from a different modality and the objective is to minimise the reconstruction loss between the input and the output using a differentiable sinkhorn operator.

PCRL (Zhou et al., 2021a) and PCRLv2 (Zhou et al., 2023) use a combination of three pretext tasks, rotation prediction (Gidaris et al., 2018), context prediction (Pathak et al., 2016), and instance discrimination (Chen et al., 2020a) for learning representation from MR scans. PCLRv2 extends PCRL to multi-scale resolutions for better performance along with other architectural changes. $SSL2$ (Wang et al., 2023a), CSwin (Li et al., 2023) also uses a similar framework for sclerosis segmentation and prostate cancer detection and segmentation, respectively. In a recent work Monsefi et al. (2024), the slices are clustered to encode different features, and a classification task is used as the pretext task, where the network predicts the suitable cluster for a collection of slices from MR scans.

(b) Instance Discrimination-based Frameworks

The work Jamaludin et al. (2017) can be regarded as one of the first applications of SSL to medical image analysis. This work uses a patient discrimination task using contrastive loss with vertebrate level prediction as an auxiliary task for representation learning. CADx (Chen et al., 2022) uses InfoNCE loss on texture information extracted from cervical optical CT images to learn representations to detect high-risk diseases, including high-grade squamous intraepithelial lesions and cervical cancer. You et al. (2021) uses a momentum based instance discrimination, dimension contrastive and consistency loss between teacher and student network along with the supervised loss to learn representations from CT scans for volumetric segmentation. Wu et al. (2021) and Wu et al. (2022b) constructs local positives and negatives from partitioned scan volumes, also takes scan partitions from different remote patients as negatives, in a federated environment, and uses a momentum-based contrastive learning framework (Chen et al., 2020c) to learn representations for volumetric segmentation in the downstream task. DrasCLR (Yu et al., 2024) uses N-pair contrastive loss by sampling positive and negative samples from the neighbourhood in addition to InfoNCE loss to learn representations of 3D lung CT images.

Chaitanya et al. (2020) used a dense contrastive representation learning framework for segmentation from MR images. OS2 (Yang et al., 2023b) uses a contrastive learning framework with a novel support query interactive embedding module (SQIE), equipped with channel-wise co-attention, spatial-wise co-attention, and spatial bias transformation blocks to mine interactive information between slices. Vox2Vec (Goncharov et al., 2023) also uses a contrastive learning framework on multi-scale representations to capture both global semantics and local semantics.

Nguyen et al. (2023) uses SwAV (Caron et al., 2020) as the baseline framework for clustering semantic representations. The dependence between 2D slices in 3D volumes is learnt by mapping the aggregated embedding from all the slices close to embeddings of individual slices. Masked embedding predictions are also used as an auxiliary task.

Windsor et al. (2021) uses contrastive learning-based dense correspondence matching between DXA and MRI scans, along with unsupervised image registration to transfer segmentation annotations between the two modalities.

MsVRL (Zheng et al., 2022) uses BYOL (Grill et al., 2020) as the baseline framework and extends it to multiscale representations from MR scans. BT-UNet (Punn and Agarwal, 2022) uses Barlow Twins (Zbontar et al., 2021) to train the encoder, which is later fine-tuned for segmentation tasks on MR scans. This work also presents performance on histopathological and skin lesions data.

(c) Non-Contrastive Frameworks

Ouyang et al. (2020), Ouyang et al. (2022) uses superpixels-based semantic segments as pseudo-labels for few-shot segmentation without fine-tuning. Li et al. (2021b) used a pre-trained network for feature extraction and subsequent k-means clustering for sample re-weighting or imbalance-aware selection. The authors use SimSiam (Chen and He, 2020) as a baseline framework for representation learning.

2.3.2 Ultrasound

There are several pieces of work which have applied SSL principles on Ultrasound data. In the subsections below, we classify the works based on the different frameworks like context-based, instance discrimination-based, etc.

(a) Context-based Frameworks

Jiao et al. (2020) uses transformation prediction and video frame order prediction as a joint prediction task to learn representation from ultrasound videos. Hu et al. (2020) combined context encoding pretext task like Pathak et al. (2016) with adversarial training and DICOM metadata prediction to form the pre-training framework.Jiao et al. (2020) attempts to correlate audio with visual features in ultrasound video by optimizing a cross-modal contrastive loss.

Qi et al. (2020) uses a jigsaw based task for learning representation from ultrasound images for utero-placental interface detection in the downstream task. Zhang et al. (2024b) uses a combination of two context-based pretext tasks, namely, rotation prediction and image pixel ordering prediction to learn representations, which are then used in the downstream fusion architecture for carotid plaque ultrasound image classification. Fang et al. (2023) also uses rotation prediction along with self-distillation based objective as the pretext task for endometrial disease classification. The rotation prediction pretext task has also been adopted in Roop et al. (2023) for estimating the angular offset of freehand ultrasound probe movement relative to an ideal viewing angle. Another contemporary work, Xie et al. (2024) uses jigsaw puzzle solving as a pretext task to learn representations from thyroid ultrasound images.

Lin et al. (2022) uses a masked video modelling framework based on TimeSFormer auto-encoder architecture (Bertasius et al., 2021) for pre-training, followed by a correlation-aware contrastive framework to enhance feature resemblance for the downstream classification task. In recent work, Fan et al. (2024) uses masked autoencoders to learn representations from breast ultrasound images for tumour classification. A similar framework is also used in Xu et al. (2024) for representation learning from tongue and breast ultrasound images. Sang et al. (2023) also uses a masked image modelling-based framework to learn representations from thyroid and breast ultrasound data for segmentation downstream tasks. SimICL (Zhou et al., 2024) uses visual in-context learning by predicting the query mask from paired query image and support image and mask as reference using a masked image modelling framework.

FetusMap (Yang et al., 2019) uses a reconstruction-based framework with a landmark detector for fetal pose estimation, aided by generating the pseudo-labels from an atlas of poses. FetusMapV2 (Chen et al., 2024) further enhances the fetal pose estimation by proposing a better memory management framework along with a pair loss to mitigate confusion caused by symmetrical and similar anatomical structures. Alasmawi et al. (2024) proposes a self-supervised representation learning step followed by a self-labelling step (Van Gansbeke et al., 2020) to cluster fetal ultrasound images. Lamoureux et al. (2023) also uses a reconstruction-based framework for cardiac ultrasound images. In Kang et al.

(2023), the authors use a deblurring-based reconstruction framework for thyroid ultrasound classification. Image registration is used as the pretext task in Ding et al. (2024) to learn representations from carotid ultrasound images for segmentation of plaque in the downstream task.

(b) Instance Discrimination-based Frameworks

Jiao et al. (2020) attempt to correlate audio with visual features in ultrasound video, along with ensuring that features of audio and video lie close to each other by minimizing a cross-modal contrastive loss. In Perek et al. (2021), a similar multi-modal contrastive framework is adopted for learning representations from mammography and ultrasound data. A multi-modal contrastive learning framework is also used in Jiao et al. (2023), where the authors used video-audio correspondence prediction and cross-modal contrastive learning framework to learn ultrasound video and speech-audio representations. The representations learnt in the pretext task are used to localise anatomical regions of interest during ultrasound imaging, with only speech audio as a reference in the downstream task.

Liang et al. (2023) uses a self-supervised contrastive framework for learning representations from ultrasound data which are then used to pseudo-label unlabelled data in the subsequent semi-supervised learning stage for anatomy tracking.

Basu et al. (2022) uses both cross-video and intra-video negative frames from ultrasound frames to learn representations. The temporal difference between the anchor and intra-video negatives is also gradually decreased to increase the hardness of the task. HiCo (Zhang et al., 2023a) uses features from multiple hierarchical levels of the encoder to implement fine, medium and coarse-grained contrastive learning in addition to global-local contrast framework to improve classification performance on lung and breast ultrasound data. Wang et al. (2024) uses transverse and longitudinal views of thyroid ultrasound data to perform single-view and multi-view contrastive learning based on MoCov2 (Chen et al., 2020c) framework but with independent encoders and a shared memory bank in the pre-training for thyroid nodule classification and segmentation in the downstream task. DSMT-Net (Li et al., 2024) also uses MoCov2 in addition to another masked image modelling branch for learning representations from endoscopic ultrasound images for detection of pancreatic and breast tumours. DSMT-net also uses a multi-operator transformation module to extract and transform the ROIs ultrasound image into rectangular input.

(c) Non-Contrastive Frameworks

Inspired by BYOL, SelfCSL (Nguyen and Le, 2021) uses the same to pre-train a randomly initialized backbone for semi-supervised learning on small-scale MedMNIST dataset (Yang et al., 2021; Yang et al., 2023a).

Abdi et al. (2024) uses the Barlow Twins framework as the baseline for learning representations to improve keypoint detection performance in Transmitral Doppler imaging, which is a type of ultrasound imaging. VanBerlo et al. (2024) uses a weight for invariance term in VICReg or Barlow Twins depending on the temporal or spatial distance between the

samples in a positive pair in lung ultrasound videos. Similarly, To et al. (2024) also uses VICReg as the pre-training framework for a prototype-based out-of-distribution detection in the downstream task. Lu et al. (2023) uses SimSiam (Chen and He, 2020) as the baseline framework to learn representations using a position and channel-based dual attention architecture from prostate ultrasound images for cancer screening.

2.3.3 Endoscopic Visual Data

In this section, we present several pieces of work which have used different SSL frameworks for representation learning from Endoscopic images or videos. In the following subsections, we discuss several such works and also classify based on the different types of framework used.

(a) Context-based Frameworks

Ross et al. (2017) uses an adversarial training strategy, where the generator produces recolorized images using a U-Net architecture, which is used for segmentation in the downstream task. Vats et al. (2021) uses rotation prediction and jigsaw puzzle-solving tasks for self-supervised pre-training from wireless capsule endoscopic images. This work also discusses the primary reasons behind the gaps that occur in the learning of semantic representation due to inadequate self-supervised training. In Hong et al. (2021), a reconstruction-based framework is used on colorectal images to learn representations for polyp segmentation.

(b) Instance Discrimination-based Frameworks

Jian et al. (2021) uses instance discrimination on endoscopic images to learn representations for the detection of Helicobacter Pylori infection. In Intrator et al. (2023), the authors primarily explore two methods, single frame instance discrimination, and multi-view tracklet discrimination. In Colo-SCRL (Chen et al., 2023b), the authors combined VideoMAE (Tong et al., 2022) with VideoMoCo (Pan et al., 2021) for representation learning from paired colonoscopy videos.

(c) Non-contrastive Frameworks

FPSiam (Gan et al., 2023) uses SimSiam (Chen and He, 2020) as the baseline framework to learn representation from frames extracted from colorectal videos. In addition to the baseline framework, FPSiam utilizes features from intermediate encoder layers to implement local feature similarity to reduce the aliasing effect of upsampling.

2.3.4 X-Ray / Radiographs

In this section, we discuss the pieces of work which have applied different SSL frameworks on X-ray or radiography images, and also classify them based on the type of SSL framework used.

(a) Instance discrimination-based Frameworks

Works like Sowrirajan et al. (2021); Chen et al. (2021d) use MoCo as a baseline framework for self-supervised pre-training. MedAug (Vu et al., 2021) uses a unique approach of using patient metadata to pair scans to construct positive pairs in contrastive learning. Tiu et al. (2022) uses a multimodal contrastive framework to learn representations from chest radiograph images, to predict pathology in the downstream task. Manna et al. (2021a) demonstrate that self-supervised pre-training of ImageNet-pretrained network on unlabeled domain-specific medical images, significantly improves the accuracy of medical image classifiers. Similar findings were also reported in MoCo-CXR Sowrirajan et al. (2021) and Azizi et al. (2021).Sun et al. (2021) uses both patch or node based contrastive learning and graph level contrastive learning to learn both global and local representations from chest radiographs. DiRA (Haghighi et al., 2022) and DiRAv2 (Haghighi et al., 2024) use a combination of image restoration, adversarial, and instance discrimination framework for learning representation from chest radiographs. ConVirt (Zhang et al., 2022c) uses paired chest radiographs and text reports for text-guided cross-modal contrastive learning of visual representations. Zhang et al. (2023c) uses a disease classifier by distilling knowledge from a network trained using cross-modal contrastive loss using paired image and text information.

(b) Context-based Frameworks

IDEAL (Mahapatra et al., 2021) takes a saliency map-based interpretability-driven sample selection approach, with only the use of an autoencoder for clustering the X-ray images using the latent feature vectors as the self-supervised part. DiRA (Haghighi et al., 2022) and DiRAv2 (Haghighi et al., 2024) also fall under this category of frameworks.

2.3.5 Retinal Images

Like the previous medical imaging modalities, there are several pieces of work which have applied different SSL frameworks on retinal image data. In this section, we discuss such pieces of work and also classify them according to the type of SSL framework used.

(a) Instance Discrimination-based Frameworks

Mojab et al. (2020) uses data from multiple devices/domains and applies SimCLR (Chen et al., 2020a) as the baseline framework to learn representations and show that a multi-domain self-supervised contrastive learning approach performs better than supervised

transfer learning. Gupta et al. (2023) also uses instance discrimination for representation learning from fundus images. Li et al. (2021b) uses two different pretext tasks but in a collaborative learning or multitasking setting. This work uses rotation prediction and patient/instance discrimination together for pre-training.

(b) Context-based Frameworks

Holmberg et al. (2020) used the macular thickness obtained from the automatic segmentation of the optical coherence tomography volume as pseudo-labels for the pretext task of predicting macular thickness from IR fundus images. The pre-trained network is then used for classification of diabetic retinopathy in colour fundus images.

Hervella et al. (2018) uses multimodal reconstruction as a self-supervised pretext task. This work is used in Álvaro S. Hervella et al. (2020) to deal with label scarcity. In Álvaro S. Hervella et al. (2021), the pretext task of multimodal reconstruction of fluorescence angiography from retinography is approached using aligned retinography-angiography pairs as pre-training data. In Hervella et al. (2020), the same pretext is used for joint optical disc and cup segmentation in images of the eye fundus.

Uni4Eye (Cai et al., 2022) proposes a masked image modelling approach with a novel unified patch embedding module to learn unified representations from 2D colour fundus images or Fundus Fluorescein Angiography (FFA), 3D optical coherence tomography (OCT) and optical coherence tomography (OCTA) images.

Yang et al. (2023) uses multi-modal masked relational modelling, to enrich the semantic relation among diseases. The relation matching is proposed to capture abundant disease-related relations by aligning sample-wise feature relations between intact and masked features in both self- and cross-modality levels.

2.3.6 Histopathology

In the literature there are several pieces of work where SSL frameworks have been applied to histopathological image data. In this section, we discuss such works and also categorize them based on the type of SSL framework.

(a) Context-based Frameworks

Štepec and Skočaj (2020) uses generative image synthesis as a pretext task for anomaly detection in the downstream task. StarDist (Prakash et al., 2020) uses a denoising framework for learning representation from biomedical microscopy images for downstream segmentation tasks. Stacke et al. (2020) uses the framework of CPC (van den Oord et al., 2018) on histopathological images for representation learning. The study found that only low-level CPC features are relevant for tumour classification.

(b) Instance Discrimination-based Frameworks

Ciga et al. (2022) applies instance-based contrastive learning on histopathological images using SimCLR (Chen et al., 2020a) as the baseline framework. DSMIL (Li et al., 2020) uses a pre-trained SimCLR backbone for weakly supervised multi-instance learning on whole slide images. Saillard et al. (2021) uses a multiple-instance learning framework on patches extracted from background subtracted whole slide images. RNAPath (Cisternino et al., 2023) uses a multi-instance learning framework too, to learn representations from 1.7M wide slide images across 23 healthy tissues in 838 donors using a ViT encoder for downstream tasks like tissue substructure segmentation. The model estimates the gene expression at the patch level by independent gene-wise linear regressions, to obtain patch-level scores, which are averaged to obtain a sample-level prediction. CELLULOSE (Wolf et al., 2023) uses a patch-based object-centric contrastive approach to allow the segmentation of individual cells in microscopy images.

2.3.7 Echocardiogram

In this section, we discuss several pieces of work where different types of SSL framework have been used for representation learning from echocardiogram visual data. However, in this section, we mostly discuss those pieces of work that have used context-based frameworks for learning representations.

Echo-SyncNet (Dezaki et al., 2021) uses multiview echocardiogram videos to learn spatiotemporal information by optimizing consecutive frame similarity, correspondence matching, and temporal order of frames. EP (Chen et al., 2021a) uses a reconstruction-based framework to synthesize ECG panorama which allows real-time querying of any ECG views from a single input view. Mehari and Strodthoff (2022) uses BYOL (Grill et al., 2020) to learn the representations of the ECG data.

Yang et al. (2022) uses a combined image reconstruction and colourization-based self-supervised learning framework to learn representations from colour Doppler echocardiography images. Lee et al. (2023) uses ConvNextv2-based masked autoencoder to learn representations to diagnose myocardial diseases such as left ventricular hypertrophy and hypertensive heart disease. SimLVSeg (Maani et al., 2024) uses a masked video modelling framework to learn representations from echocardiogram videos for downstream left ventricle segmentation tasks. In a recent work Damasceno et al. (2024), the authors use DINOv2 (Oquab et al., 2023) based pre-training framework to learn representations from Transthoracic Echocardiography to detect Pulmonary hypertension.

2.3.8 Skin Images

Another medical imaging modality which has also found its use in self-supervised representation learning is skim image data. In this section, we discuss several pieces of work which have applied SSL frameworks on skin image data, and also categorize them based on the type of SSL framework used.

(a) Context-based Frameworks

JIANet (Zhang et al., 2022a) uses a jigsaw shuffled skin lesion image as one sample in a positive pair in a jigsaw invariant instance discrimination task. This work also uses a VAE-based reconstruction branch as part of the proposed collaborative learning framework. The reconstruction branch serves as the means to preserve the important semantic features necessary for melanoma segmentation in the downstream task.

Zhi et al. (2024) uses both a masked autoencoder and a self-distillation based framework to learn representations from skin images. The student network in the self-distillation framework shares parameters with the encoder in the masked autoencoder branch. To enhance generalization, Zhi et al. (2024) applies exterior conversion augmentation and dynamic feature generation to the inputs to the teacher network.

(b) Instance Discrimination-based Frameworks

STCN (Wang et al., 2021a) uses a combination of transformation invariance, reconstruction, and pseudo-label based classification tasks for learning representations. The pseudo-labels are obtained by clustering the embeddings obtained from the encoder using a modularity-inspired deep topology clustering algorithm.

Wang et al. (2023) uses dermoscopy and clinical skin images for multi-modal contrastive learning. Yang et al. (2024) uses a combination of discriminative and generative SSL for skin lesion classification.

MHC-PO (Liang et al., 2023) combines self-supervised contrastive learning with supervised classification tasks. However, to adjust the conflicting gradients between contrastive clustering and classification, the authors use a pareto optimization phase based on the CAGrad algorithm, a multi-objective gradient manipulation method.

(c) Non-Contrastive Learning-based Frameworks

In one of the most recent works, BOLT (Li et al., 2024) combines ViT-based BYOL with a difficulty awareness loss. As a pre-processing step, BOLT perturbs the input tokens. An auxiliary task predicts which branch, student or teacher is processing the tokens with a larger level of perturbation, and is optimized by the difficulty awareness loss. Useini et al. (2024) uses a self-distillation framework DINO as the baseline, whereas Morita and Han (2023) uses a BYOL-based network with an adaptive augmentation module for efficient self-supervised representation learning.

2.4 Conclusion

In this chapter, we discussed several studies on self-supervised learning conducted over a considerably long time frame and gained several insights about the research trends. We

observed that a large number of studies has been done in the last few years on masked image modelling. Among the paired embedding-based frameworks, SimCLR, MoCov2, BYOL, and SimSiam are the most popular and they are adopted for several downstream applications. We also discussed some of their derivative works which have laid the path for adoption and subsequent application of the same in different domains. The works are also categorized based on the principle of their respective baseline framework, which allows a deeper understanding of these works on both conceptual and fundamental levels. The analysis of the characteristics of different types of frameworks presented in this chapter allows the identification of the differences between each type of framework.

We then discussed the works implementing self-supervised learning frameworks for representation learning from medical imaging modalities like magnetic resonance, computed tomography, ultrasound, retinal images, histopathology images, etc. Please note that, in Section 2.3, although different types of SSL methods are discussed based on the types of medical images to get a clear idea of the wide applicability across the different types of medical imaging modalities, we didn't consider all types of medical images in our thesis for experimentation and hence we did not address all types here. It can be observed that the modalities with the higher number of applications of SSL are the magnetic resonance and computer tomography modalities, followed by ultrasound and histopathology images. This observation can be attributed to the relatively abundant availability of magnetic resonance or computed tomography data compared to the other modalities. Through the discussions in this chapter, we observe the wide spectrum of pretext tasks that have been discovered and effectively implemented for representation learning from unlabeled data.

In the next chapter, we adopt the jigsaw puzzle-solving strategy for context-based representation learning from medical images. The jigsaw puzzle-solving strategy serves as a tool to learn object composition as well as the constituent parts, thereby learning the spatial relations between the different regions in a medical image. However, jigsaw puzzle-solving-based pretext tasks are not immune to issues like shortcut learning or redundant representations. Furthermore, it is essential to learn context-invariant representations for better performance in classification-based downstream tasks. To this end, we propose a framework to learn context-invariant representations and also diminish the effect of redundant representations in the pre-training or pretext stage.

Chapter 3

Context-Based Self-Supervised Learning for Medical Image Analysis

3.1 Introduction

Computer Vision tasks like object detection, segmentation, or tracking have reached near human precision and reliability with increased labelled data and scaling of computational resources for supervised learning frameworks. Furthermore, with the invention of better architectures like ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021), etc. the need for a huge amount of annotated data for learning proper generalization has also increased. In the absence of sufficient data, supervised deep learning models often suffer from the overfitting problem. In the case of medical image data, collection requires specialized apparatus, and annotation needs to be done with the help of experienced medical personnel. The availability of experienced medical personnel is also scarce. Thus, collecting and annotating such huge amounts of data is expensive and time-consuming. Moreover, medical image datasets often suffer from the class-imbalance problem, which requires special attention (Huynh et al., 2022) as it is often difficult to find an equal number of people with a specific diagnosis and collect their data with the same configuration of tools.

The drawbacks of the supervised learning algorithms as mentioned above have led researchers to devise techniques that enable models to learn meaningful representations of the data without training from scratch. One such technique is called *Transfer learning*. However, in medical image analysis, researchers often use ImageNet-pretrained weights on datasets which differ a lot from the distribution of ImageNet (Deng et al., 2009). This disparity in data distributions often leads to difficulty in fine-tuning and can also destroy the hierarchical co-adaptation among the features (Yosinski et al., 2014). Furthermore, if the scale of the target/downstream medical data is small, then there is a risk of over-fitting as well. Thus, it is required that the distribution of the data on which the pre-trained network is trained should be similar to the data used in the downstream task. Such endeavour has led to the advent of techniques which can learn representation from data without the need

for human-annotated labels. Such techniques fall under the paradigm of self-supervised learning.

Advances in context-based self-supervised learning techniques (Gidaris et al., 2018; Jing and Tian, 2018; Feng et al., 2019; Pathak et al., 2016; Doersch et al., 2015) which can learn meaningful representations of the underlying data without human annotations were some of the initial attempts. The features learned by the pretext model serve as the *pre-trained* features in the downstream task and hence, the pretext tasks are carefully designed such that the pretext model can extract meaningful representations from the data. In Gidaris et al. (2018), representations are learned by predicting the angle by which an image has been rotated. In Pathak et al. (2016), the model is made to generate the missing contents from the image. This requires the model to understand the surroundings of the missing part and learn the visual semantic structures of the data as well. Among other techniques, image colouring (Zhang et al., 2016) maps the grayscale input to quantized colour value in the CIE *Lab* colour scale and treats the problem as a multinomial classification. Methods like Noroozi and Favaro (2016); Wei et al. (2019); Ahsan et al. (2019) use jigsaw puzzle solving as the pretext task. In this chapter, the frameworks discussed are also based on this strategy. Noroozi and Favaro (2016) argued that jigsaw puzzle-solving can serve as a valuable teaching tool for systems to learn about object composition and its constituent parts. Although the mapping of individual tiles in a jigsaw puzzle to specific object parts may be uncertain, observing all tiles collectively can resolve these ambiguities due to the mutually exclusive nature of tile placement.

In the aforementioned works, the representations learned by the model are based on natural images. Despite the development of several self-supervised learning algorithms for natural images or videos, the application of such algorithms in medical image analysis remains limited. Medical data needs to be handled in a different way than natural image data because of its nature. Medical data like magnetic resonance imaging or computed tomography images are grayscale. While the region of interest may or may not cover a substantial spatial area, the temporal presence of the same may have limited presence. Furthermore, the dimensions of medical data are often large because of the need to capture minute details for proper diagnosis. Thus, a model needs to be capable of learning fine-grained representations from the data. In our initial investigation, we applied context-based methods like geometric transformation prediction by discretizing the transformation parameters into classes to predict. In self-supervised learning methods like rotation prediction (Gidaris et al., 2018) or geometric transformation prediction (Jing and Tian, 2018), the boundary pixels are relocated to new positions leaving voids in their original locations. Because of this, the boundary pixels having higher intensity give rise to the formation of some low-level features. These features, e.g. blank areas, image boundaries, image corner points, etc. in the transformed slices of MR scans help the pretext model optimize the objective function without learning useful representations. Although geometric transformation prediction works well for natural images (Gidaris et al., 2018; Jing and Tian, 2018), the behaviour was not the same in our experiments. This led to the learning of shortcut solutions which resulted in the network learning representations which were not useful for the downstream task. However, learning context-invariant representations for better performance in the classification-based downstream tasks is essential.

However, even jigsaw puzzle-solving tasks are also not free from the evil effects of shortcut learning and need special architectural design, as can be observed from Sec. 3.3.2. To this end, in this chapter, we introduce two frameworks based on jigsaw puzzle solving. The proposed novel deep learning frameworks for self-supervised representation learning for knee injury diagnosis, based on jigsaw puzzle solving aim to learn useful representations along with preventing the learning of low-level and redundant representations or shortcut solutions in the pretext or pre-training stage. The pre-trained weights are used in the subsequent downstream stage for the classification of injuries in the Knee MR scans. As the objective of the pretext stage is to learn representations from the data using pseudo-labels, it is expected that the presence of an imbalance in the ground truth classes affects the representation learning. In our work, the downstream task is a classification problem and its objective is to predict one or more anomalies present in the MR scans.

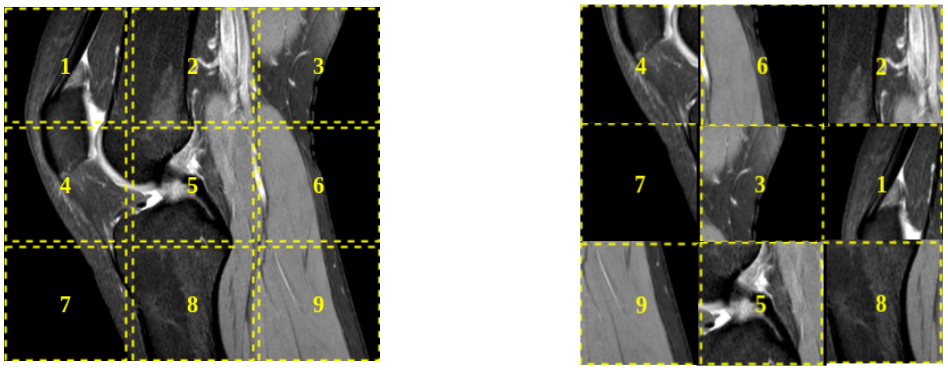
While the first framework laid the foundation for representation learning from MR scans using jigsaw puzzle solving by using a semi-parallel architecture to prevent learning of low level signals as depicted in Sec. 3.3.2, the improved second framework aims at better representation learning, and points at several intriguing aspects of self-supervised learning, such as the role of the number of model parameters in the pretext phase. From hereon, we will refer to the first framework as SSACL as it deals with only ACL tear detection in the downstream task and the second framework as SKID as it deals with a complete diagnosis of knee injury from MR scans.

The rest of the chapter is organized as follows: In Sec. 3.2, we present the preliminary information regarding the jigsaw puzzle problem, and hamming distance which are used in our proposed jigsaw puzzle-solving frameworks. In Sec. 3.3, we discuss the motivation which drives the proposed strategies. Next, in Sec. 3.4, we describe the proposed pretext tasks along with patch arrangement strategy and channel-wise aggregation strategy to decouple the representations from each patch and prevent learning of redundant representations. In Sec. 3.5, we provide the experimental details, present comparative results and analyses, and also discuss ablation studies. Lastly, we conclude this chapter in Sec. 3.6.

3.2 Preliminaries

Jigsaw Puzzle Solving

In the context of computer vision or images specifically, the problem of jigsaw puzzle solving can be described as rearranging the patches in an image. Suppose, we have an image I of dimensions $H \times W$. The first step involves dividing the image I into 9 patches, each with dimensions $H' \times W'$, such that $3 \times H' \leq H$ and $3 \times W' \leq W$. The position of each patch is denoted by numbers from 1 to 9 in a zig-zag pattern, such that the patch on the top left corner is numbered 1 and the one on the bottom right corner is numbered 9. One such sample is shown in Fig. 3.1.a. After jumbling the patches if the patches are put together, it will look like the image shown in Fig. 3.1.b.



(a) (b)

FIGURE 3.1: (a) Example of an image showing the position of the patches in order. (b) Image showing the patches after being arranged in a randomly chosen order.

Hamming Distance

The hamming distance between two strings of equal length is equal to the number of places where the two strings are different. In other words, the minimum number of substitutions required to transform one string into another string of the same length is termed as hamming distance. It satisfies the conditions of non-negativity and symmetry and also follows the triangle inequality (Cover and Thomas, 2006).

In the context of jigsaw puzzle solving, we can represent the arrangement of the patches as shown in Fig. 3.1 as a string. For instance, the arrangement of the patches in Fig. 3.1.a can be written as 123456789, and the arrangement of the patches in Fig. 3.1.b can be written as 462731958. The hamming distance between the two arrangements is equal to 9.

3.3 Motivation

In this section, we will discuss the motivation behind the framework proposed in this chapter. We discuss two-fold motivation, firstly, the failure of geometric transformation prediction related to context-based pretext tasks to avoid learning shortcut solutions, consequently harming downstream performance. Secondly, the failure of single-stream convolutional architecture to avoid shortcut solutions as well. The above two cases are discussed in the following sections. Keeping these failure cases in mind, we primarily design an architecture, which employs a separate stream or branch for each of the 9 patches. This prevents the learning of low-level signals and artefacts. Furthermore, it also emphasises the learning of context invariant representation by randomizing patches over all the branches.

3.3.1 Shortcomings of Geometric Transformation Prediction Task

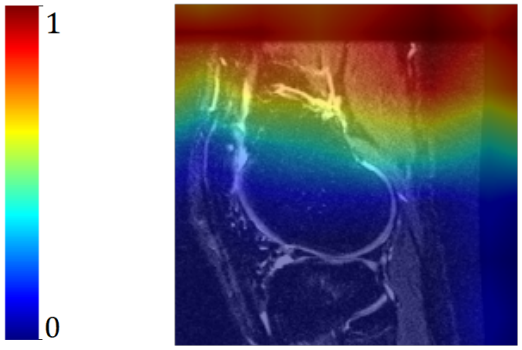
Visual biomedical data like ultrasound videos, magnetic resonance scans, and X-ray images are essentially grayscale images. The features of interest are only a few pixels wide. Besides, these features are also present only for a few slices in spatiotemporal biomedical data. Thus, learning these features from biomedical data requires very intricate and fine-grained learning. In addition to the temporal prevalence of the features, the spatial distribution of the pixels plays an important role in representation learning. As the slices in videos/scans of biomedical data are grayscale, the boundary transition regions from high to low-intensity pixels pose problems in feature learning.

To support the above findings on Geometric Transformation Prediction, we modelled a classification task and chose the transformation from a finite set G, which can be expressed as a Cartesian product of four finite sets $R, \mathcal{T}r_x, \mathcal{T}r_y$ and S, i.e. $G = R \times \mathcal{T}r_x \times \mathcal{T}r_y \times S$ where $R = \{-15°, 0, 15°\}$, $\mathcal{T}r_x = \{-\lfloor 0.1H \rfloor, 0, \lfloor 0.1H \rfloor\}$, $\mathcal{T}r_y = \{-\lfloor 0.1W \rfloor, 0, \lfloor 0.1W \rfloor\}$ and $S = \{1, 1.2\}$. Here $R, \mathcal{T}r_x, \mathcal{T}r_y$ and S denote the finite sets of angles of rotation in degrees, magnitude of translation along the x-axis and y-axis in pixels and scale factors, respectively. H and W denotes the height and width of the slices of the MR scans. The total number of combinations of geometric transformations is 54. The pretext model used for testing the credibility of the geometric transformation prediction task consisted of a pre-trained VGG Net with 54 outputs as we have used a combination of 54 transformations in total. The images in Fig. 3.2 show the gradient class activation mappings (Selvaraju et al., 2017) on a few slices from each plane (Sagittal, Coronal, and Axial) after a random transformation is applied. It can be observed that the regions of interest in the GradCAM (Selvaraju et al., 2017) outputs are mostly edges, empty regions, or corners, which are low-level features. Thus the model utilises the boundary discontinuities to classify the geometric transformations in the grayscale medical images. In other words, the model learns a shortcut path to obtain a global minimum in the loss landscape. Choosing the proper pretext task is the most important part and the results shown in Fig. 3.2 have greatly influenced our work.

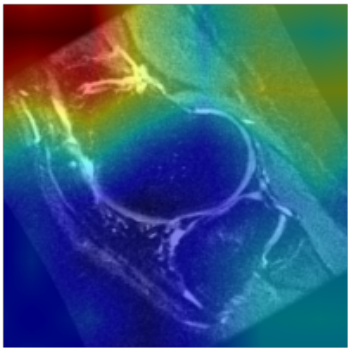
The learning process and the decision taken by the downstream model in the inference stage depend on the features extracted by the pretext model, as we already mentioned in Sec. 3.5.2.2 that the pretext model was kept frozen in the downstream task. Thus, the gradient class activation mappings obtained from the last Dimension Reduction block of the pretext model are a direct reflection of the features extracted by the pretext model. In the gradient class activation mappings in Fig. 3.10, we can see that the regions of interest are important in making the correct decision in inference. Whereas, in Sec. 3.3.1, we can observe from Fig. 3.2, that the regions of interest are not aligned with any features of interest in the MR slices, which explains why the standalone geometric transformation prediction task failed to learn any meaningful representations.

3.3.2 Shortcomings of Non-Parallel Architecture

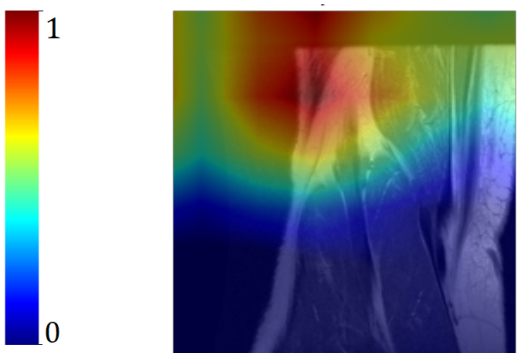
Pretext task models are very prone to learning low-level signals like void regions, boundary edges and corners, etc. Without any strategy to prevent inter-patch information leakage,



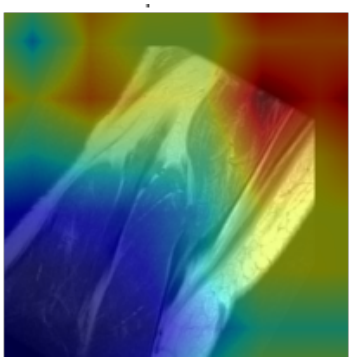
(a) Translation



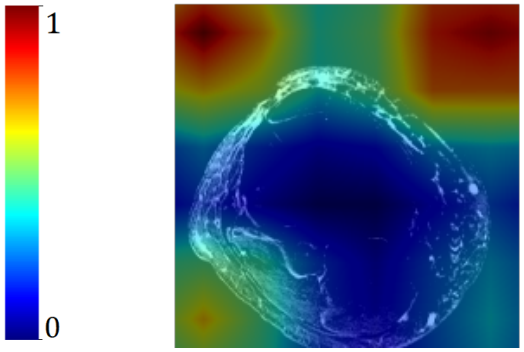
(b) Rotation



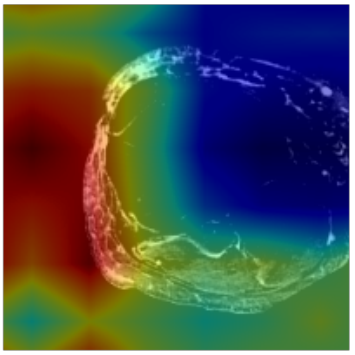
(c) Translation



(d) Both Translation and Rotation



(e) Translation



(f) Both Translation and Rotation

FIGURE 3.2: Gradient class activation mappings of slices from 3 planes, showing the regions of interest in a Geometric Transformation Prediction task (2 instances from each plane are shown). (a) & (b), (c) & (d), and (e) & (f) belong to Sagittal, Coronal, and Axial planes, respectively. The individual captions indicate the geometrical transformation applied to the slices.

the model tends to learn low level signals similar to the clues that humans often use when solving jigsaw puzzles. The approach we follow in this chapter also compels us to take a subset of the large pool of possible rearrangements.

We used an Inception-ResNet-v2 (Szegedy et al., 2017) network pre-trained on ImageNet, to predict the jigsaw arrangements. The input was all the 9 patches put together like in Fig. 3.1.b. After analysing the gradient class activation mapping (Selvaraju et al., 2017)

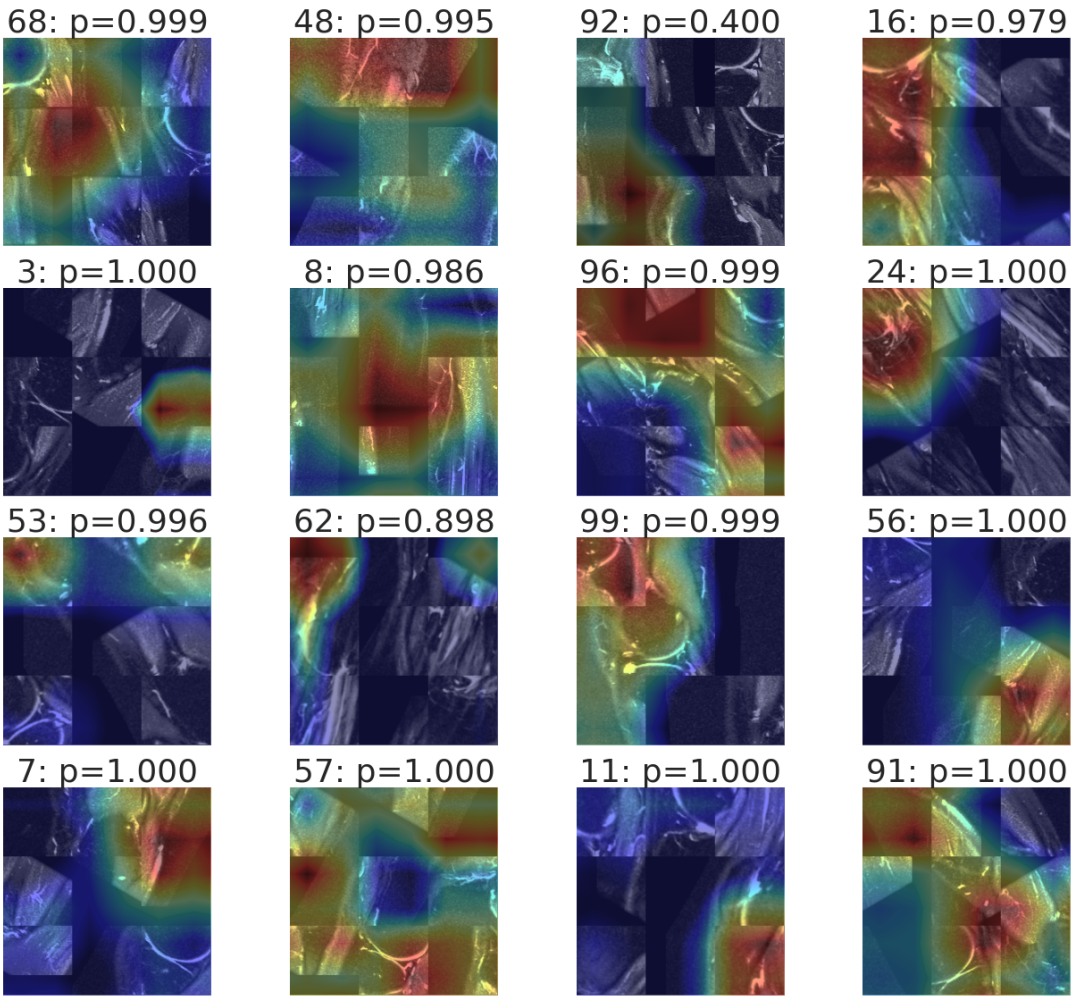


FIGURE 3.3: Gradcam output shows the regions (indicated by red) where the model built using pre-trained Inception-ResNet-v2 gains maximum information. It is clearly visible that the maximum attention is on the low-level signals as mentioned in Section 3.3.2

outputs of the aforementioned model as given in Fig. 3.3, we observed that the model used the low-level signals like boundary corners, edges and discontinuities between patches as discriminative features. This tendency of the model to learn features without proper generalization of the loss surface prevents it from learning meaningful context-invariant visual representational features.

3.4 Proposed Framework

In this section, we discuss the proposed framework, which consists of two stages, namely, (1) Pretext task framework, and (2) Downstream task framework. As we discuss two different methods in this chapter, the corresponding framework for each of those two methods are discussed sequentially.

3.4.1 Pretext Task Frameworks

In this subsection, we will discuss the pretext task algorithm followed for both our frameworks. The pretext task is designed as a multi-class classification task. In this learning framework, a slice $\mathcal{F}$ is chosen randomly according to a uniform distribution from a magnetic resonance scan. This slice is then divided into N_P square patches. Each patch is of dimensions $\lfloor \frac{H}{\sqrt{N_P}} \rfloor \times \lfloor \frac{H}{\sqrt{N_P}} \rfloor$, where H is the length of each side of each square slice $\mathcal{F}$, that is, $H = W$. Dividing a slice into N_P patches gives $N_P!$ ways to jumble the N_P patches. If we consider $N_P = 9$, the number of ways the patches can be jumbled is $9! = 3,62,880$.

3.4.1.1 Patch Arrangement Selection Strategy

Let, $\mathcal{J}$ be the set of all possible arrangements of the patches. Solving a classification task with such a large number of classes would require huge computational time and resources. For the SSLACL framework, we choose a subset $\mathcal{A} \subset \mathcal{J}$ by following the Algorithm 3.1 (SETARR) which describes the steps used to choose the permuted rearrangement orders to be included in $\mathcal{A}$. We initialize the set of arrangements with the ordered arrangement $[1, 2, 3, 4, 5, 6, 7, 8, 9](\tau_0)$ and choose the threshold of hamming distance as $\lfloor \frac{N_P}{2} \rfloor$. The hamming distance between two permutations is defined as the number of positions in which they differ. In our experiments, we consider $N_P = 9$, thus the threshold of hamming distance equates to 4. We progressively keep on adding elements from the set $\mathcal{J}$ if the hamming distance from all the elements in the set $\mathcal{A}'$ is more than 4. This algorithm ensures that the elements in the chosen set are neither too close nor too far from other elements in the permutation space. This maintains a balance in the difficulty of the pretext classification task. However, for the SKID framework, we simply choose 1000 elements from the set $\mathcal{J}$ randomly according to a uniform distribution and form set A.

Algorithm 3.1: SETARR : Set of arrangements selection

Result: $\mathcal{A}$: Set of Arrangements
Given
$\mathcal{J}$: Set of all possible arrangements
$\mathcal{U}_A$: A is a sample drawn from uniform distribution $\mathcal{U}$
$\mathcal{C}$: Number of classes in the pretext task
Initialize $\mathcal{A}' = \mathcal{A} = \{ \tau_0 \}$;
for *i = 1 : 9! - 1* **do**
  if $hammingDist(a, \mathcal{J}[i]) > 4 \; \forall \; a \in \mathcal{A}'$ **then**
    $\mathcal{A}' = \mathcal{A}' \bigcup \{\mathcal{J}[i]\}$;
  end
end
for *i = 1 : $\mathcal{C}$ - 1* **do**
  $\mathcal{A} = \mathcal{A} \bigcup \mathcal{U}_A[\mathcal{A}']$
end

3.4.1.2 Jumbled Patch Generation Strategy

To obtain the jumbled patches and the pretext labels, we apply the algorithm PREPFRAM (Algorithm 3.2) to the selected slices. First, a slice is divided into $\sqrt{N_P} \times \sqrt{N_P}$ parts with each part having dimensions $\lfloor \frac{H}{\sqrt{N_P}} \rfloor \times \lfloor \frac{H}{\sqrt{N_P}} \rfloor$ as shown in Fig. 3.1.a. Augmentation g is applied to each of the N_P patches to generate the set of augmented patches P_g. Then, for each patch P' in P_g, we sample a point (ref_x, ref_y) from the range $[0, \lfloor \frac{H}{\sqrt{N_P}} \rfloor - 64]$ randomly according to a uniform distribution $\mathcal{U}$. Using this point (ref_x, ref_y) as origin, we then crop a 64×64 region (P'_{64}) from the patch. All the cropped patches (P'_{64}) comprise the set of patches $P_{g,64}$. In our experiments, we set $\lfloor \frac{H}{\sqrt{N_P}} \rfloor = 85$. Finally, the arrangement $\mathcal{A}_\tau$ is applied on the patches $P_{g,64}$ to get the jumbled patches P_A as shown in Fig. 3.1.b.

Algorithm 3.2: JUMPAT: Jumbled Patch Generation

Result: P_A : Jumbled Patches from a slice $\mathcal{F}$
Given
A : Set of rearrangements
G : Set of geometric transformations
$\mathcal{U}_s[\cdot]$: s is a sample drawn from uniform distribution $\mathcal{U}$ over any set
$\mathcal{T}$: an arrangement to be applied on the patches and sampled uniformly from set A
$\mathbf{map}_I(\cdot)$: a function which denotes a transformation being applied to an instance
Initialize
$\mathcal{F}$: a random slice from an MR scan sample
$P_A, P_{g,64} = \{\,\}$
$H' = \lfloor \frac{H}{\sqrt{N_P}} \rfloor$
$row = col = ref_x = ref_y = 0$
for *i=1:9* **do**
  $row = \lfloor \frac{i}{\sqrt{N_P}} \rfloor$
  $col = i \mod \sqrt{N_P}$
  $P' = \mathcal{F}[row.H' : (row+1)H', col.H' : (col+1)H']$
  $g = \mathcal{U}_g[G]$
  $P' = \mathbf{map}_I(g, P')$
  $ref_x = \mathcal{U}_x[0, H' - 64]$
  $ref_y = \mathcal{U}_y[0, H' - 64]$
  $P'_{64} = P'[ref_x : ref_x + 64, ref_y : ref_y + 64]$
  $P_{g,64} = P_{g,64} \bigcup P'_{64}$
end
$\mathcal{T} : \mathcal{U}_\tau[\mathcal{A}]$
$P_A = \mathbf{map}_I(\mathcal{T}, P_{g,64})$

3.4.1.3 Data Augmentation Strategy

Data augmentation was applied to increase the diversity of the data and make the model robust. The data augmentations applied include horizontal and vertical shift, rotation, and scaling. Horizontal and vertical shift values were randomly selected from $[-\lfloor 0.1H_p \rfloor, \lfloor 0.1H_p \rfloor]$ and rotation values were selected from the range $[-15°, 15°]$. The scale factor was however chosen from a finite set $S = \{1, 1.2\}$. Here, H_p denotes the length of each side of a square patch and $H_p = 64$ in our experiments. We chose the augmentation randomly from a finite

set G, which can be expressed as a Cartesian product of four finite sets R, $\mathcal{T}r_x$, $\mathcal{T}r_y$ and S. We also applied additive white Gaussian noise with mean 0 and variance 0.01 to the image patches before feeding them to the model.

3.4.1.4 Semi-Parallel Convolutional Architecture

In this chapter, we have used a semi-parallel convolutional architecture for our pretext tasks to predict the order in which the patches are arranged. The architecture is shown in Fig. 3.4. The results presented in the previous section show the reason behind the adoption of a semi-parallel architecture in this chapter. We feed each of the 9 patches in the input into one of the 9 parallel convolutional channels. Each convolutional channel is made up of 2 Convolutional blocks. Each convolutional block consists of two convolutional layers followed by a maxpooling layer. The number of filters of both the convolutional layers, in the two convolutional blocks are 256 and 512, respectively. The maxpooling layer has a pool window of dimensions 2×2 and a stride of 2. We can check whether the semi-parallel part of the architecture justifies the title of this subsection in the results presented in Sec. 3.5.3.4.

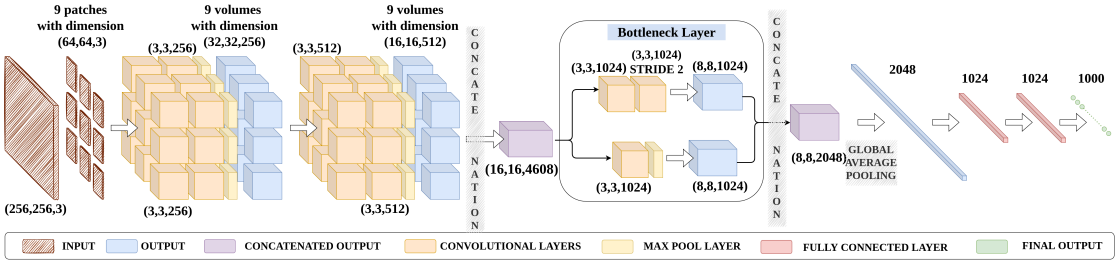


FIGURE 3.4: Proposed network model for pretext task in the SSLACL framework.

3.4.1.5 Channel-wise Feature Aggregation

In the SSLACL framework, we followed a channel-wise feature aggregation step. Having the information from all the patches concatenated along the channel dimension allows the network to process complete information from all the patches at each spatial location. According to our intuition, this prevents context-dependent representation learning. The output from all the 9 channels is then concatenated to get an output volume of dimension $16 \times 16 \times 4608$. This output volume is then convoluted with a convolutional layer with filters $3 \times 3 \times 2048$ to reduce the dimensionality and gives an output of dimensions $16 \times 16 \times 2048$, which is then fed into the bottleneck layer as shown in Fig. 3.4. It consists of two separate branches. The first branch is a convolutional block, which consists of two convolutional layers with 1024 filters, with only the second layer having stride 2, thereby causing the spatial dimension of the output to be reduced to half of its input. The second branch contains a convolution layer with 1024 filters with kernel size 3×3 and followed by a maxpooling layer which reduces the dimensions to half. The outputs from the two channels are again concatenated to form an output volume of dimension $8 \times 8 \times 2048$. Global average pooling is applied to this output of dimension $8 \times 8 \times 2048$ to obtain an output of dimension

2048, which is then fed into a network of two fully connected layers of dimension 1024. The second fully connected layer is connected to the output consisting of $\mathcal{C}$ nodes, where $\mathcal{C}$ is the number of classes.

Similarly, in the SKID framework, we followed the same channel-wise aggregation strategy to ensure that the network learns context-invariant representations in the pretext tasks. However, to enhance the quality of representations learnt in the pretext phase, we made some additional modifications in the SKID framework, discussed in the section below. However, in the SKID framework, the number of channels in the second convolutional block was reduced to 256 from 512 in the SSLACL framework to reduce the computational overhead of the subsequent pipeline in SKID.

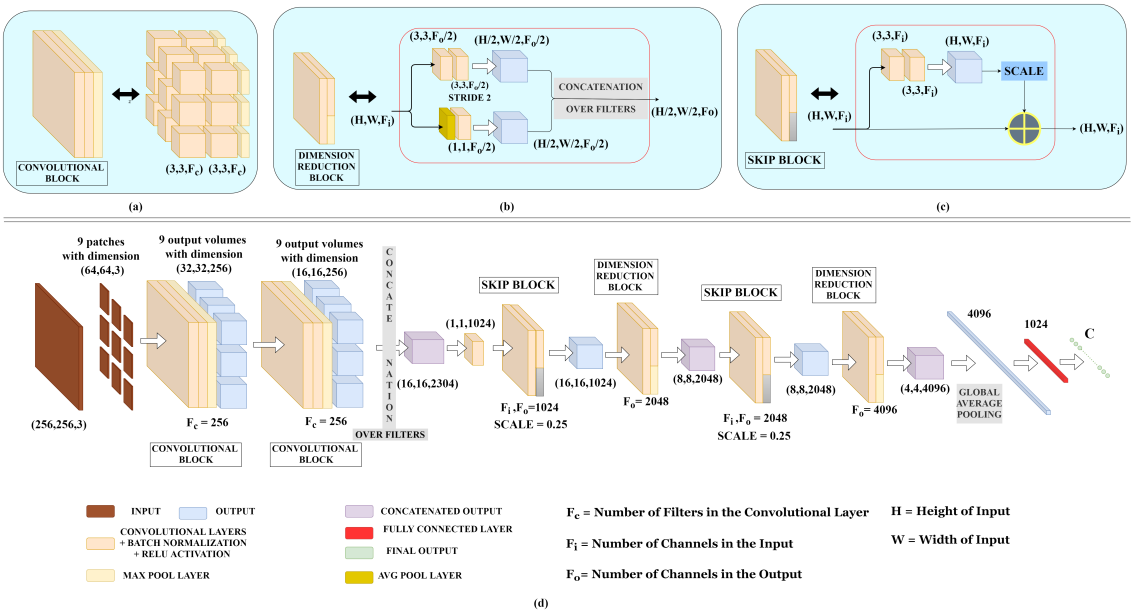


FIGURE 3.5: Proposed model for pretext task in the SKID framework. The model architecture contains three types of blocks: (a) Convolutional, (b) Dimension Reduction and (c) Skip. These three blocks are presented in an expanded view above the model diagram and marked as (a), (b), and (c). The pretext model architecture is shown in (d).

3.4.1.6 Dimension Reduction and Skip Blocks

The outputs obtained from the 9 convolutional branches are concatenated to form a resultant output of shape $16 \times 16 \times 2304$. This output is fed to a 1×1 convolutional layer with 1024 filters to reduce the dimensionality of the output. There are two main types of blocks in our architecture, *Skip* block and *Dimension Reduction* block. The architectures of both blocks are shown in Fig. 3.5. The output from the *Skip* block is the summation of its input and the scaled output from two successive convolutional layers with filters of dimension 3×3. The scale factor in the *Skip* block influences the model's performance. In the *Dimension Reduction* block, we combine strided convolutions and average pooling to reduce the spatial dimension.

The *Dimension Reduction* block consists of two branches, as shown in Fig. 3.5. The upper branch consists of two convolutional layers, with the first and second layers having strides 1 and 2, respectively. The lower branch consists of an average pooling layer followed by a 1×1 convolutional layer. The number of filters in all the convolutional layers in the *Dimension Reduction* block is the same as the number of channels in the input. The outputs from the two branches are concatenated to obtain the final output with reduced spatial dimensions. The architecture of our *Pretext* model consists of two of each such block, with a *Dimension Reduction* block succeeding every *Skip* block, as shown in the Fig. 3.5. The output of the second *Dimension Reduction* block is of the shape $4 \times 4 \times 4096$. Global Average Pooling is applied to this output to obtain an output of dimension 4096, which is then fed to a fully connected layer with 1024 nodes. Finally, another fully connected layer with $\mathcal{C}$ nodes is connected to the last layer, where $\mathcal{C}$ is the number of classes in the classification task. Depending on the number of input and output channels in the *Skip* and *Dimension Reduction* blocks, we have 3 variants of the model. The details of the above-mentioned 3 variants of the model are presented in Table 3.1.

The Skip block is inspired by ResNet (He et al., 2016) skip connections. In spatiotemporal data like MR data, the extent of occurrence of some desired feature is short both spatially and temporally. This requires the model to capture the intricate features to learn representations that can yield performance comparable to supervised learning algorithms in downstream tasks. As the extent of the desired features is short, the Skip blocks allow the model to learn the essential features by utilizing the residual signals. The skip connections allow the model to prevent vanishing gradient problems and also utilize the features from the initial layers in the higher layers for feature learning. The Dimension Reduction block aims at learning the downsampling procedure and thereby preventing information loss. This block also contains an average pooling layer, which is not learnable. Thus, this block combines a fixed operation with a learnable operation to capture more responsive features along with increased expressibility. The Dimension Reduction block also serves as a bottleneck layer as the model is trained to learn the downsampling operation, which trains it to learn useful representations and discard redundant information. The effect of adding the Skip block and the Dimension Reduction block is evident from Tables 3.5 and 3.6. in Sec. 3.5.3.1, where we present the performance of our model with modifications of the Skip and the Dimension Reduction block (Proposed-NoBlocks), to show the functions of these blocks. In the model named "SKIDv3-NoBlocks", we replaced the Skip block with an identity function $f(x) = x$, and the Dimension Reduction Block was replaced by an average pooling layer with a window of dimensions 2×2. We observe that the performance dropped considerably for the modified model. The performance of the different variations of the pretext model is presented in Sec. 3.5.4.3.

3.4.2 Downstream Task Frameworks

We approach the downstream task in SSLACL and SKID in different ways. In SSLACL, the downstream task is simply a binary classification task to predict whether the Knee MR scans of the Sagittal modality contain Anterior Cruciate Ligament (ACL) injury. To this end, we choose only to utilise the representations learnt by the convolutional blocks in the downstream task. This necessitates the use of a task-specific classifier to learn task-specific high-level representations in the downstream task. Therefore, the downstream

TABLE 3.1: Different Variants of the Pretext model. 'Conv.' refers to the term 'Convolutional', and 'params' refers to the word 'parameters'.

SKID Model variant	Number of Filters in								No. of params
	Conv. Block	1×1 Conv. layer	1^{st} Skip block		1^{st} Dim. Red. Block Output	2^{nd} Skip block		2^{nd} Dim. Red. Block Output	
			1^{st} Conv. layer	Output		1^{st} Conv. layer	Output		
SKID-v1	256	1024	512	1024	1024	512	1024	4096	$\approx$ 109 M
SKID-v2	256	1024	512	1024	2048	1024	2048	4096	$\approx$ 170 M
SKID-v3 (Proposed)	256	1024	1024	1024	2048	2048	2048	4096	$\approx$ 217 M

model consists of two parts: the feature extractor and the classifier, as shown in Fig. 3.6. From the pretext model, the 9 branches with 4 convolutional layers each act as the feature extractors. The outputs from the 9 branches are concatenated to form an output of dimensions $(64 \times 64 \times 512)$ and fed to the classifier.

3.4.2.1 SSLACL Divide-and-Teach Strategy

We also devise a unique *Divide-and-Teach* training strategy. Since the slices were uniformly sampled from each MR scan, each convolutional layer can extract useful features from the slices, irrespective of the sequential/temporal position of the slice in the MR scan. We divide the N_S slices into 9 parts before feeding to the 9 channels of the CNN, N_S being the total number of slices in the MR scan, following Algorithm 3.3. After the respective outputs are obtained from each channel, we concatenate the outputs over the slices to obtain an output of dimension $N_S \times 64 \times 64 \times 512$, which is then fed into the classifier to obtain the predictions.

Algorithm 3.3: DIVFRAM: Dividing Slices in an MR scan

Result: I : Input to the downstream network
Initialize
I = [ ];
start_index = **end_index** = 0;
Given
F : All slices in the MR scan
N_S : Number of slices
for *i = 1 : 9* **do**
  $n = \lceil \frac{N_S}{9-i+1} \rceil$
  end_index = end_index + n
  Append **F** [start_index : end_index] to **I**
  start_index = end_index
end

3.4.2.2 SSLACL Downstream Model Architecture

The classifier consists of three convolutional blocks, each containing two convolutional layers. Both the layers in each convolutional block have filter size 3×3 but only the second convolutional layer has stride 2. This reduces the dimensions to half without the use of maxpooling layers. Using strided convolutions also allows the models to learn the downsampling process and prevents information loss. The three convolutional layers result in an output of shape $N_S \times 8 \times 8 \times 1024$. We then apply Global Average Pooling to the output, followed by maxpooling over slices following MRNet (Bien et al., 2018). This gives an output of dimension 1024, which is then fed into a network of two fully connected layers, each containing 1024 nodes. The output from this layer is finally fed into the output node. The downstream task is a binary classification task, hence sigmoid activation is applied on the output node to obtain predictions in the 0 to 1 range.

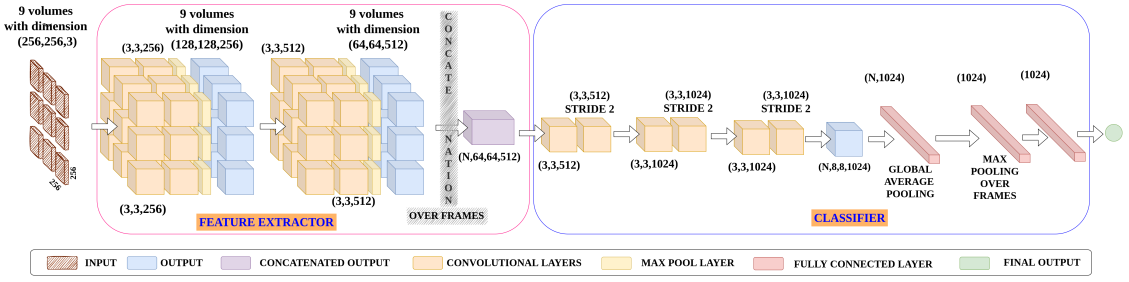


FIGURE 3.6: Proposed network model used for the downstream task in the SSLACL framework.

In the SKID framework, however, the downstream task is an imbalanced multi-label classification task. In SKID, we attempt to classify the Knee MR scans for three different abnormalities: Anterior Cruciate Ligament tear, Meniscus tear, and Abnormality which includes the previous two as well. We model the downstream task as 3 separate binary classification tasks for 3 classes, following the Binary Relevance method (Godbole and Sarawagi, 2004). We trained a different model for each of the three different planes and then used an ensemble of the models for the final prediction as described in the following subsections.

3.4.2.3 SKID Ensembling Strategy

In this SKID framework, we use the ensemble of the three models trained on three planes of MR data: Sagittal, Coronal, and Axial, to obtain the final results during inference on the validation set. The main reason for doing so, instead of training a single model on all three planes, is that the nature of the images in the three planes is different. This causes the distribution of the images from the three planes to differ spatially. The details of the ensembling strategy used for inference in the downstream task are given below.

We used the weighted majority voting (Zhou, 2012) approach for the ensemble. It is essential to give more weight to the stronger classifier for each class. The output prediction

of the ensemble of the 3 models is given by

$$H_j(x) = \sum_{i=1}^{T} w_i^j f_i^j(x) \tag{3.1}$$

where w_i^j is the weight assigned to the classifier output $f_i^j(x)$ for the j^{th} class and i^{th} classifier, and N_C is the total number of classifiers. Here, N_C is equal to 3, as we use a separate classifier for each of the three planes. The weights are non-negative and are constrained by $\sum_{i=1}^{N_C} w_i^j = 1$. The weights are calculated as

$$w_i^j = \ln(\frac{acc_i^j}{1 - acc_i^j}) \tag{3.2}$$

where acc_i^j is the prediction accuracy for the j^{th} class by the i^{th} classifier.
From the gradient activation class mappings obtained after training the downstream models on the three planes - Sagittal, Coronal, and Axial (shown in Fig. 3.10), we found that the features contributing more towards the final predictions are distributed normally along the time axis. Based on this information, for evaluation of the downstream model, we sampled the slices of the MR scan from a normal distribution with $\mu = \frac{N_S}{2}$ and $\sigma = \frac{N_S}{4}$, where N_S is the number of slices in the MR scan. In addition to the sampling strategy, we follow a Monte Carlo method to infer the final predictions on the validation set. We sampled 16 slices 8 times from each scan and obtained 8 predictions for each scan. The final prediction for that particular scan was obtained by averaging those 8 predictions, that is,

$$f_i^j(x) = \frac{1}{8} \sum_{s=1}^{8} f_i^{j,s}(x) \tag{3.3}$$

where f_i^j is final prediction, and $f_i^{j,s}$ is the classifier prediction output for the s^{th} sample of 16 slices. The final predicted label is obtained by,

$$\hat{y}_j = H_j(x) > 0.5 \tag{3.4}$$

where $\hat{y}_j$ is the predicted label for the j^{th} class.

3.4.2.4 SKID Downstream Model Architecture

The downstream model architecture (Fig. 3.7) consists of the pretext model as the feature extractor and a Convolutional LSTM (ConvLSTM) network (Shi et al., 2015) as the classifier for efficient handling of the temporal/sequential correlation present in each scan, instead of the maxpooling over slices operation in SSLACL or MRNet (Bien et al., 2018). The downstream classifier consists of two ConvLSTM layers each with 512 output channels. The slices from the MR scans are fed into the pretext model and the output obtained

from the second *Dimension Reduction* block is fed into the classifier. The 4D output tensor from the feature extractor is reshaped and passed to the ConvLSTM network. Global Average Pooling is applied to the output obtained from the ConvLSTM network to get an output feature vector of 512 dimensions. The final layer consists of 3 nodes for 3 classes with a sigmoid activation function. The total number of parameters in the downstream model is approximately 103 Million.

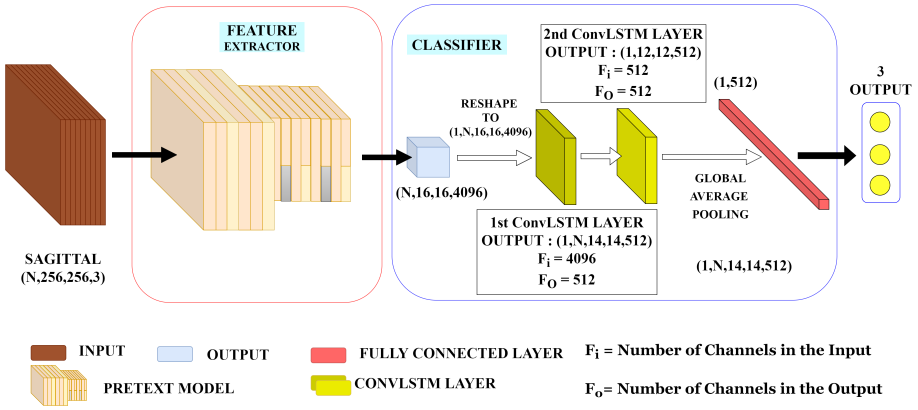


FIGURE 3.7: Network model used for the downstream task in the SKID framework.

3.5 Experimental Details, Results and Analysis

In this section, we discuss the datasets used in our experiments, the implementation details of the pretext and downstream tasks, and the results obtained in the downstream experiments followed by a comparison with the supervised baseline MRNet (Bien et al., 2018).

3.5.1 Datasets

To validate the proposed self-supervised framework we use the benchmark MRNet (Bien et al., 2018) dataset, which contains 1370 scans. Out of the 1370 scans, 1130 scans are used as the training set and 120 scans are considered as tuning or validation set. The remaining 120 scans are used for testing the model. The classes in the dataset are Abnormality, ACL Tear, and Meniscus Tear. Out of the 1,130 training samples, 917 are Abnormal exams, 208 are ACL tears and 397 samples are Meniscus tears. The MRNet dataset is an imbalanced multi-label dataset. The multi-label and imbalanced nature of the dataset can be observed in the concurrence plot in Fig. 3.8. This allows us to explore the effects of self-supervised learning techniques on a class-imbalanced multi-label classification task. The average number of slices in each scan is 30.4.

We also used the KneeMRI (Štajduhar et al., 2017) dataset for external validation by fine-tuning a model pre-trained on the MRNet dataset. The dataset consists of 917 12-bit grayscale volumes of either left or right knees. Each volume record was labelled according to the ligament condition: (1) healthy, (2) partially injured, or (3) completely ruptured.

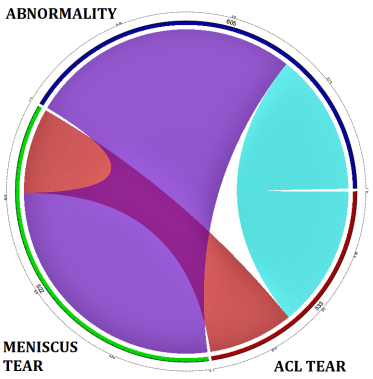


FIGURE 3.8: Concurrence plot of the labels in the MRNet dataset showing the interactions between different labels.

The dataset contains 706 healthy, 174 partially injured, and 55 completely ruptured samples. We use the dataset in two types of tasks: (1) binary classification task with the healthy and the partially injured samples as the negative samples and the completely ruptured samples as the positive samples, as mentioned in (Bien et al., 2018), (2) imbalanced multi-class classification setting. The training, validation, and test set splits used in the experiments are the same as. The different folders in the KneeMRI dataset, used for the training, validation and test set splits are given below (here the folder names are mentioned as 'vol01-10'):

TABLE 3.2: Data splits of KneeMRI dataset (Štajduhar et al., 2017).

Train	Validation	Test
'vol08', 'vol04', 'vol03', 'vol09', 'vol06', 'vol07'	'vol10', 'vol05'	'vol01', 'vol02'

FIGURE 3.9: Region of Interest for ACL tear detection. Images (c) and (d) show the enlarged view of the ROIs marked in red in the image (a) and (b), respectively. The images in Fig. 5.a (also 5.c) and 5.b (also 5.d) are examples of a torn ACL and an uninjured ACL, respectively.

3.5.2 Implementation Details

In this section, we discuss both the pretext and downstream implementation details for SSLACL and SKID frameworks.

3.5.2.1 Pretext Implementation Details

During training, the model was trained with slices chosen randomly according to a uniform distribution from MR scans. We optimized the categorical cross-entropy loss of the model using RMSProp optimizer with a learning rate of 10^{-4} and exponential decay of 0.95 every epoch. We used a batch size of 32 for SSLACL and 16 for SKID during both the training and validation stages. The batch size was decreased for SKID as the number of model parameters increased. However, during the validation stage, we did not use any augmentations to the input. The SSLACL pretext model was trained on a 12GB NVIDIA TESLA K80 GPU on Google Cloud Platform, whereas the SKID pretext model was trained on a 15GB NVIDIA Tesla T4 GPU on Google Colab. The training was stopped when the validation accuracy plateaued.

3.5.2.2 Downstream Implementation Details

The downstream tasks for SSLACL and SKID are different. Hence, the implementation is also different.

SSLACL Downstream implementation: In the SSLACL downstream task, due to memory constraints on the NVIDIA TESLA K80 GPU, we limited the number of slices to $min(N_S, 36)$, where N_S is the number of slices in an MR scan. If the number of slices in any MR scan is more than 36, we use uniform random sampling to select 36 slices from N_S number of slices. This strategy helps the model deal with missing slices and also temporally sparse data. Also, we kept the batch size limited to 1. The downstream model was trained by optimizing the binary cross-entropy loss of the model using Adam optimizer with an initial learning rate of 10^{-5}. Since the dataset is highly imbalanced, we used oversampling to balance the dataset before training our model. The number of positive and negative ACL tear injury samples in the MRNet dataset are 208 and 922, respectively. We oversampled the minority class to 922. This oversampled dataset was then used to train the downstream model. During the validation stage also, we chose $min(N_S, 36)$ number of slices and then partitioned the slices into 9 parts using Algorithm 3.3. Apart from the *Divide-and-Teach* training strategy mentioned in Sec. 3.4.2.1, data augmentations like random rotation, translation and scaling were also applied on each slice during training.

SKID Downstream implementation: SKID deals with solving a multi-label classification task. To fulfil our objectives, we froze the weights of the pretext model and only trained the classifier in the downstream model. The downstream model was trained on a 16 GB NVIDIA P100 GPU. As we also wish to understand the temporally correlated features in the MR scan and how they affect the inference, only 16 slices which is almost half of the average number of slices in an MR scan (30.4), are chosen randomly according to a uniform distribution from each scan for training the downstream model.

The downstream model with only 103 million parameters was trained by optimizing a weighted binary cross-entropy loss with a very low learning rate starting from 10^{-5} and decayed by 0.95 every epoch for a maximum of 20 epochs. The augmentations applied to the chosen slices are the same as those applied in the pretext experiments. For each plane

in the dataset, we trained a different model. An ensemble of the three models was done using the weighted majority voting approach as described in Section 3.4.2.3 to obtain the final predictions.

3.5.3 Comparative Results and Analysis

In the SSLACL framework, for the ACL tear detection task from the Sagittal KNee MR scans, the best results were obtained using our final model (Fig. 3.6) with 77 million parameters. It achieved an accuracy of 76.62% (95% CI 74.50, 78.83) on the validation set and an AUC score of 0.848 (95% CI 0.828, 0.865).

In the SKID framework, we used all three modalities, Sagittal, Coronal and Axial for predicting all three classes in the MRNet dataset, Abnormality, ACL tear, and Meniscus tear. Both binary accuracy and AUC score are reported as metrics of performance in our experiments. In Table 3.3, we present the results of the three downstream models and the ensemble of those models for each of the three classes- Abnormality, ACL Tear, and Meniscus Tear on the validation set of the MRNet dataset. During the evaluation of the validation set, we chose the slices according to the sampling strategy described in Section 3.4.2.3. To analyze the efficiency and reliability of our method, we show the gradient class activation mappings (Selvaraju et al., 2017) for the detection of all three classes in Fig. 3.10. The salient regions in Fig. 3.10 are the regions where the pretext model gains maximum information, which is then fed to the ConvLSTM network in the downstream task.

3.5.3.1 Comparison with Supervised Algorithms

To compare our methods with supervised learning techniques we present the results of the MRNet (Bien et al., 2018) model on the same dataset. In Table 3.4, we compare SSLACL and SKID frameworks on ACL tear detection tasks only with the supervised baseline MRNet. For the ACL tear detection task, apart from limiting the number of slices to a maximum of 36, the MRNet (Bien et al., 2018) was trained using the original conditions.

However, the SKID framework predicts three types of anomalies in the Knee MR scans from all three modalities, and the comparison with the supervised baseline differs in this case. The weights learnt in the pretext task serve as a good initialization point for the downstream model and result in good convergence if compared to supervised as shown in this section. The results show that even though the distribution of the features was imbalanced during the pretext task, the downstream result does not show any bias towards any particular label. This is evident from the fact that we only trained the ConvLSTM part in the downstream task and froze the parameters of the feature extractor, which was trained on the imbalanced multi-label dataset MRNet without any measures for dealing with the imbalance. As the number of positive occurrences for abnormality exceeds that of the other two labels, the model is expected to learn features mostly from the majority label. Consequently, this should affect the performance of the downstream task. However,

TABLE 3.3: Evaluation results on the validation set of MRNet dataset in the SKID downstream task.

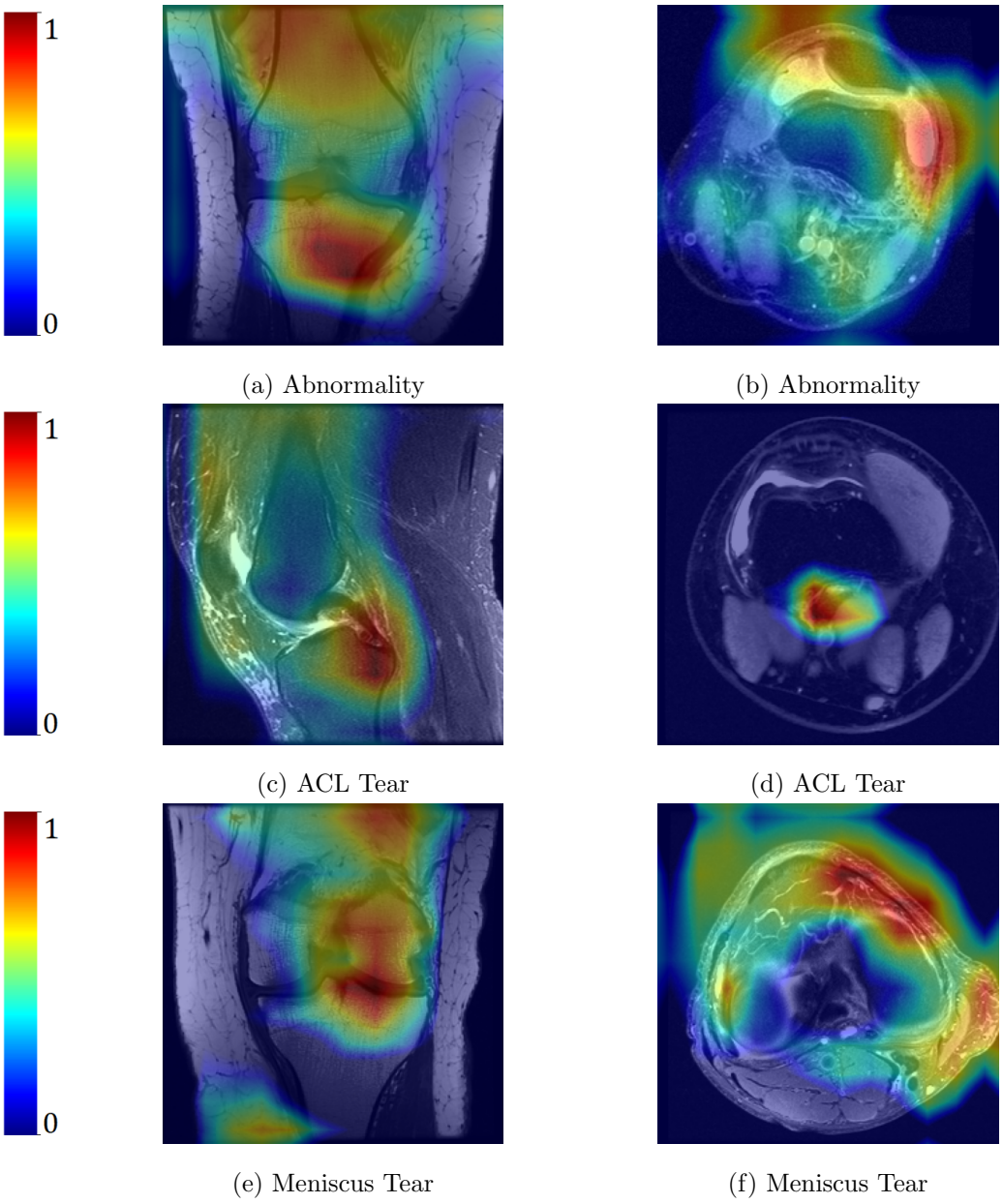
Class	Plane	Accuracy (5%-95% CI)	Sensitivity (5%-95% CI)	Specificity (5%-95% CI)	AUC (5%-95% CI)
Abnormality	Sagittal	0.883 (0.869-0.896)	0.968 (0.956-0.979)	0.555 (0.500-0.606)	0.901 (0.883-0.918)
	Coronal	0.860 (0.843-0.875)	0.957 (0.944-0.969)	0.474 (0.423-0.529)	0.847 (0.819-0.873)
	Axial	0.843 (0.829-0.856)	0.947 (0.935-0.958)	0.439 (0.375-0.500)	0.867 (0.839-0.897)
	Ensemble	**0.874** (0.862-0.887)	0.979 (0.971-0.986)	0.486 (0.432-0.543)	**0.904** (0.880-0.916)
ACL Tear	Sagittal	0.740 (0.720-0.758)	0.630 (0.597-0.665)	0.833 (0.807-0.862)	0.848 (0.828-0.867)
	Coronal	0.715 (0.695-0.738)	0.793 (0.758-0.822)	0.649 (0.613-0.682)	0.813 (0.791-0.828)
	Axial	0.807 (0.785-0.826)	0.721 (0.687-0.754)	0.879 (0.858-0.906)	0.862 (0.845-0.881)
	Ensemble	**0.800** (0.778-0.818)	0.740 (0.709-0.769)	0.849 (0.825-0.867)	**0.893** (0.878-0.909)
Meniscus Tear	Sagittal	0.653 (0.630-0.675)	0.731 (0.697-0.764)	0.587 (0.555-0.620)	0.740 (0.715-0.764)
	Coronal	0.717 (0.698-0.736)	0.981 (0.972-0.993)	0.517 (0.484-0.551)	0.803 (0.781-0.825)
	Axial	0.668 (0.649-0.689)	0.748 (0.713-0.785)	0.602 (0.564-0.634)	0.744 (0.718-0.768)
	Ensemble	**0.725** (0.706-0.746)	0.923 (0.903-0.942)	0.574 (0.539-0.608)	**0.810** (0.784-0.826)

TABLE 3.4: Comparison with supervised learning method on ACL Tear detection only.

Method	ACL Tear	
	Accuracy (%)	AUC
MRNet (Bien et al., 2018)	86.63	0.963
SSLACL	80.0 (95% CI 77.8-81.8)	0.893 (95% CI 0.878-0.909)
SKID	74.0 (95% CI 72.0-75.8)	0.848 (95% CI 0.828-0.867)

we can observe in Tables 3.5 and 3.6 in Sec. 3.5.3.1 that the performance of our model is at par with the supervised model.

For the multi-label classification task from the multi-modal MRNet data, to compare our algorithm with supervised algorithms we replicated the MRNet (Bien et al., 2018) model and trained only on 16 slices chosen randomly from the scans according to a uniform distribution, like in our downstream experiments for fair comparison. The MRNet model



(a) Abnormality (b) Abnormality

(c) ACL Tear (d) ACL Tear

(e) Meniscus Tear (f) Meniscus Tear

FIGURE 3.10: Gradient class activation mappings showing salient regions for different conditions in the Knee MR scans, obtained from the last Dimension Reduction block in the frozen pretext model of the downstream task in the SKID framework.

was evaluated on slices chosen using the same sampling strategy mentioned in Section 3.4.2.3. We present the comparison results in Table 3.5 and Table 3.6.

We can see that using only 16 slices which is almost half of the average number of slices per scan, the result of our proposed model SKID is comparable with the original MRNet (Bien et al., 2018) model. As the AUC scores are better suited as a metric for imbalanced data instead of accuracy, we can say that the representations learnt from the pretext stage of the proposed framework are not affected by the imbalance in the ground truth labels.

TABLE 3.5: AUC score comparison of SKID framework with supervised baseline MRNet. ABN: Abnormality, ACL: ACL Tear, MEN: Meniscus Tear

Method	AUC scores			
	ABN	ACL	MEN	Average
MRNet (Bien et al., 2018)	0.944	0.915	0.822	0.894
SKIDv3 (Proposed)	**0.904**	**0.893**	**0.810**	**0.869**
SKIDv3-NoBlocks	0.883	0.815	0.729	0.809

TABLE 3.6: Accuracy comparison of SKID framework with supervised baseline MRNet. ABN: Abnormality, ACL: ACL Tear, MEN: Meniscus Tear

Method	Accuracy			
	ABN	ACL	MEN	Average
MRNet (Bien et al., 2018)	0.850	0.867	0.725	0.814
SKIDv3 (Proposed)	**0.874**	**0.800**	**0.725**	**0.7997**
SKIDv3-NoBlocks	0.825	0.733	0.608	0.722

3.5.3.2 Comparison with Contrastive Learning Algorithms

In this subsection, we attempt to compare our novel task-specific architecture to some of the state-of-the-art contrastive learning algorithms. One important point worth mentioning before proceeding further is that the experiments conducted using contrastive learning algorithms were constrained by the availability of accelerator memory. The encoder in SimCLR (Chen et al., 2020a) or MoCo-v2 (Chen et al., 2020c) has a similar structure as used in the experiments in PIRL (Misra and van der Maaten, 2019) to account for the jigsaw transformation. Thus, the implementations of SimCLR Chen et al. (2020a) and MoCo-v2 (Chen et al., 2020c) deviate only in terms of the encoder structure in our experiments. The originality of the PIRL (Misra and van der Maaten, 2019) algorithm and the encoder structure was preserved to the fullest. However, the size of the memory bank had to be reduced to 1024 because of memory constraints. Apart from these, no other experimental configuration was altered for the training of the encoders in the different contrastive learning algorithms. The primary transformation in the pretext experiments was *Jigsaw* transformation, similar to PIRL (Misra and van der Maaten, 2019).

3.5.3.3 External Validation by Fine-tuning on Different Target Dataset

For external validation on both types of tasks (binary and multi-label classification), we tested the capability of the proposed model to learn transferable features by comparing the performance in the downstream task, with the pretext network trained on the MRNet (Bien et al., 2018) and the KneeMRI (Štajduhar et al., 2017) datasets. In the SSLACL framework, however, transferring the features learnt in the pretext training phase on the

TABLE 3.7: AUC Score Comparison with Contrastive Learning Algorithms.

Method	AUC Scores			
	Abnormality	ACL Tear	Meniscus Tear	Average
PIRL (Misra and van der Maaten, 2019)	0.653	0.655	0.590	0.633
SimCLR (Chen et al., 2020a)	0.676	0.691	0.663	0.677
MocoV2 (Chen et al., 2020c)	0.682	0.389	0.600	0.557
SSLACL	-	0.848	-	-
SKID	**0.904**	**0.893**	**0.810**	**0.869**

TABLE 3.8: Accuracy Score Comparison with Contrastive Learning Algorithms.

Method	Accuracy			
	Abnormality	ACL Tear	Meniscus Tear	Average
PIRL (Misra and van der Maaten, 2019)	0.750	0.600	0.567	0.639
SimCLR (Chen et al., 2020a)	0.742	0.625	0.592	0.653
MocoV2 (Chen et al., 2020c)	0.733	0.458	0.558	0.583
SSLACL	-	0.766	-	-
SKID	**0.874**	**0.800**	**0.725**	**0.7997**

MRNet (Bien et al., 2018) dataset to the downstream model yields an AUC score of 0.6799. Pretext training with the SSLACL framework on the KneeMRI (Štajduhar et al., 2017) and then transferring the trained weights to the downstream binary task yields an AUC score of 0.740.

In the multi-label classification task setting in the SKID framework, the downstream model trained with the pretext network that was trained on the MRNet dataset achieved an accuracy of 92.6% and an AUC score of 0.761 on the test split of the KneeMRI dataset. When the downstream model was trained with the pretext network that was trained on the KneeMRI dataset, the downstream model achieved an accuracy of 90.4% and an AUC score of 0.741. The difference in performance can be attributed to the difference in the number of samples in the two datasets. As the number of samples in the MRNet dataset is more than that in the KneeMRI dataset, the pretext network learnt better features, which is reflected in the downstream performance. To deal with the imbalance in the number of samples for each binary label, we resorted to oversampling the minority class. In the imbalanced multi-class classification task setting, the downstream model trained with the pretext network that was trained on MRNet and KneeMRI datasets achieved an accuracy of 70.7% and 68.6% on the test split of the dataset, respectively.

3.5.3.4 Proposed Architecture Prevents the Learning of Shortcut Solutions

In the SSLACL framework, we primarily vary the number of jigsaw arrangements to classify in the pretext task to test if the framework is capable of solving harder tasks, as using

a large number of jigsaw arrangements makes the pretext task harder. The proposed framework shows the capability of predicting the jigsaw arrangements without resorting to shortcut features.

In this subsection, we have presented the results of ACL tear injury detection from Knee MR scans. In Fig. 3.9, we can observe the region which needs to be focused on. To analyze the generalization and feature learning capacity of the model, we train with 500 and 1000 random permutations chosen according to Algorithm 3.1. As shown in Table 3.9, even after increasing the number of permutations, the proposed model performs well on the pretext tasks. Furthermore, from the GradCAM results presented in Fig. 3.11, it can be observed that unlike the GradCAM results of the geometric prediction task in Fig. 3.2, and single branch architecture in Fig. 3.3, the proposed pretext model does not rely on the low-level signals like the empty spaces, discontinuous boundaries or edges. This observation justifies the capability of the model to learn meaningful features to efficiently distinguish between such a large number of permutations of image patches.

We observe similar results with the SKID pretext model too. In Fig. 3.12, we can observe the gradient class activation mappings (Selvaraju et al., 2017) from the pretext model for all three modalities, axial, coronal and sagittal. This again proves the effectiveness of the semi-parallel pipeline in preventing the model from learning shortcut solutions or low-level signals.

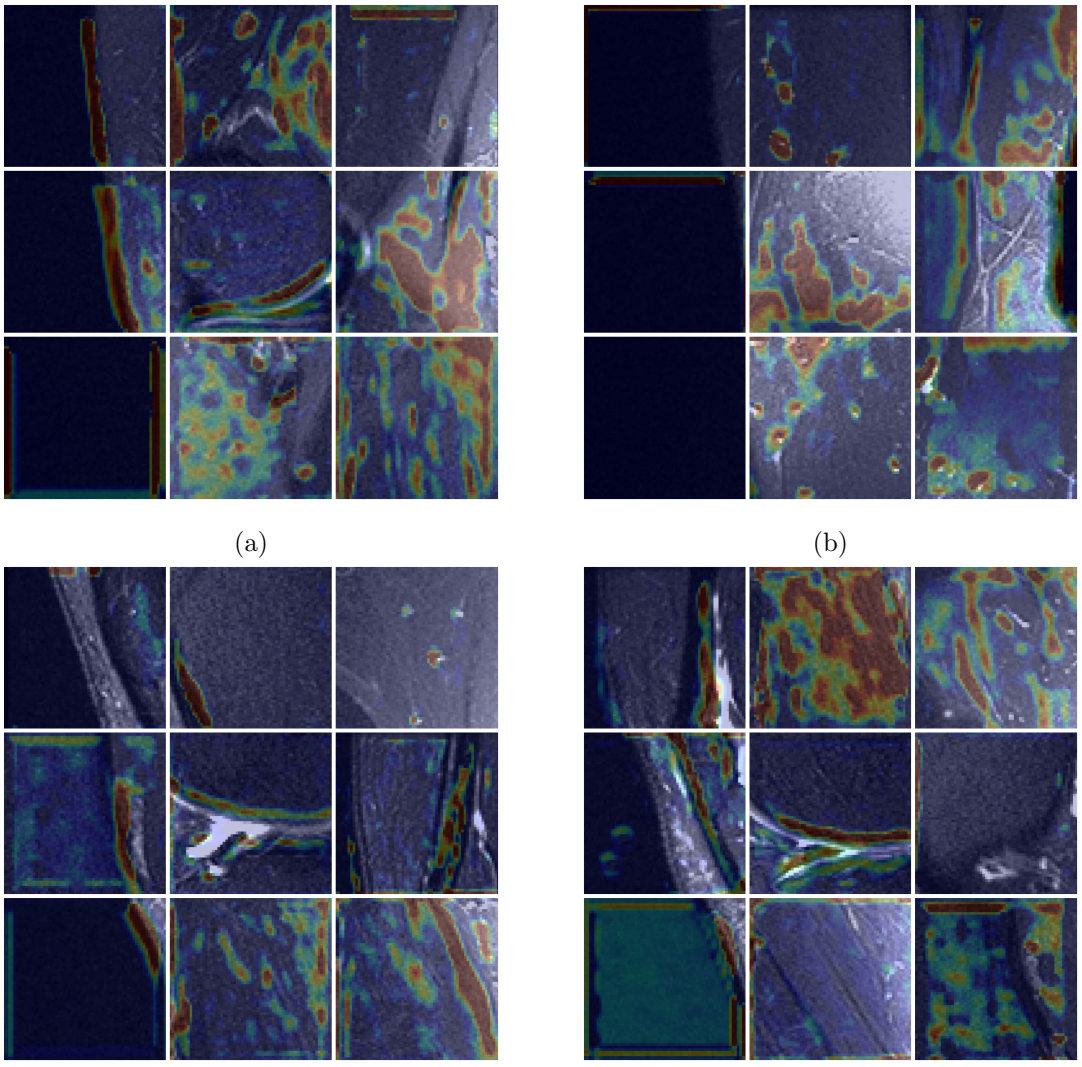
Table 3.9: SSLACL pretext task experimental results

No. of permutations	No. of parameters	Validation Accuracy
500	173 Million	96.4%
1000	173.5 Million	93.5%

3.5.4 Ablation Studies

3.5.4.1 Effect of Strided Convolution as a Downsampling Method Instead of Maxpooling Layers in SSLACL Downstream Task

To optimize our model architecture, we built multiple models by changing the different hyperparameters associated with the model. Among all the variants, the model shown in Fig. 3.6 corresponds to the final model which performed the best in the downstream task. In a variant (Model-1), a maxpooling layer was introduced instead of the first convolutional block in the classifier and the two convolutional layers in the second convolutional block contained 512 filters each. Also, only one fully connected layer was used in Model-1. In the second variation (Model-2), we increased the capacity by adding another fully connected layer with 1024 nodes and increasing the number of filters of the convolutional layers in the second convolutional block to 1024. The maxpooling layer in the classifier of Model-1 remains unchanged in Model-2. In the best performing model (as shown in Fig. 3.6), we replaced the maxpooling layer in the classifier with a convolutional block containing two convolutional layers with 512 filters each and only the second layer has a stride of 2. The performance results of all the three models have been shown in Table 3.10.



(a) (b)

(c) (d)

FIGURE 3.11: GradCAM results from the pretext task in SSLACL. The 9 patches have been rearranged according to the predicted arrangement by the SSLACL pretext model.

TABLE 3.10: Ablation study on downstream task for detection of ACL injury in SSLACL.

Model	No. of parameters	Accuracy (5%-95% CI)	AUC (5%-95% CI)
Proposed	77 Million	76.62 (74.5-78.83)	0.848 (0.828-0.865)
Model-2	75 Million	73.4 (71.0-75.6)	0.834 (0.812-0.850)
Model-1	72 Million	71.7 (70.2-72.9)	0.813 (0.797-0.829)

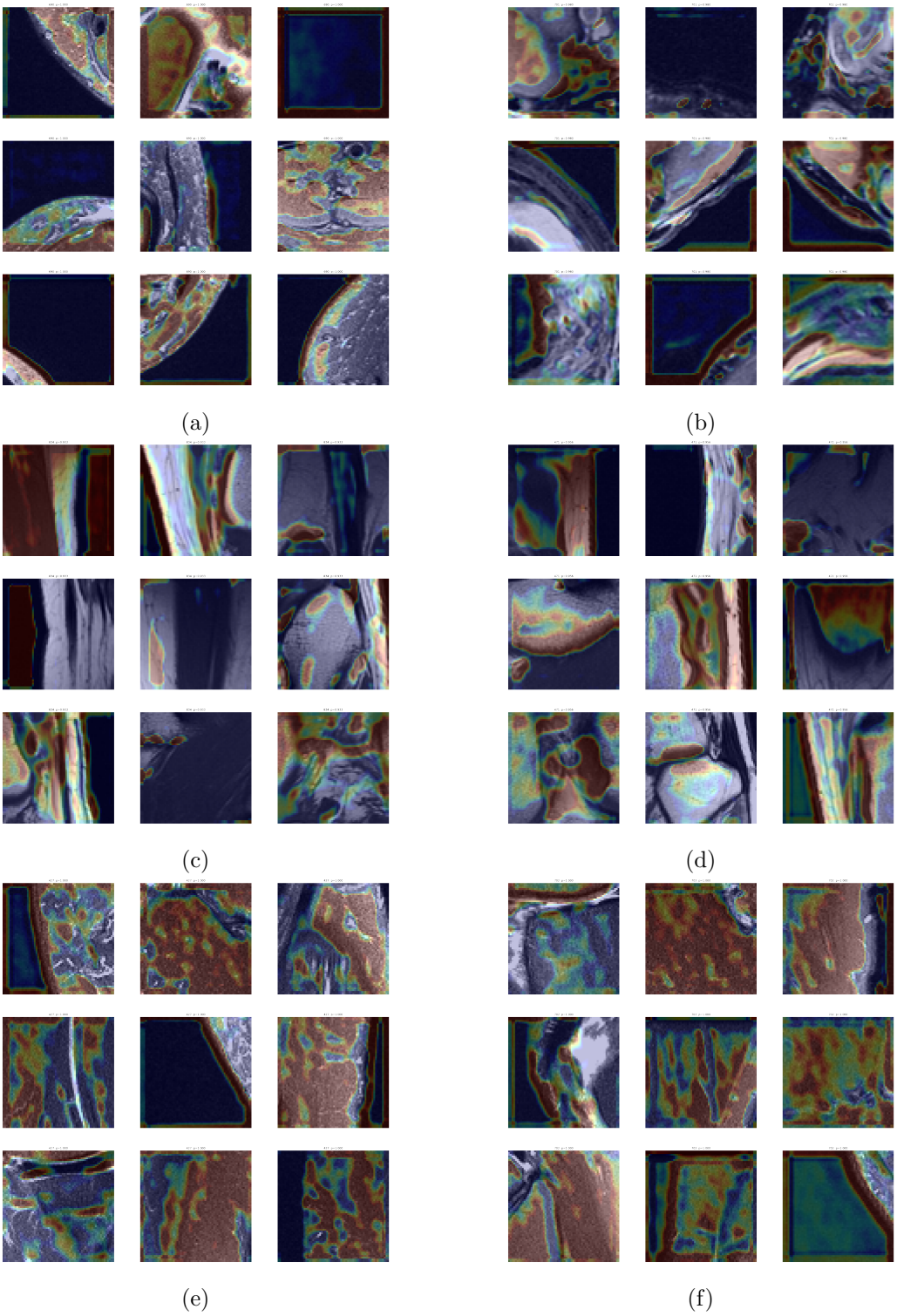


FIGURE 3.12: GradCAM results from the pretext task in SKID for all three modalities: Axial (a and b), Coronal (c and d), and Sagittal (e and f).

3.5.4.2 Effect of Data Imbalance in SSLACL Pretext Task on Downstream Performance

The pretext and the downstream task, both contribute to the ultimate objective of detection of ACL injury from knee MR scans. The motivation of our work is to build a pretext

model, capable of learning spatial context invariant visual representational features. The results presented in Table 3.11 show that in the case of an imbalanced dataset, the features of the majority class receive more weightage than the minority class in the pretext task. Every sample in the training set is chosen exactly once when preparing the pretext training samples. Thus, for each pretext label, there are more samples from the majority class than from the minority class.

When the oversampled dataset is used to train the model in the pretext task, an equal number of samples from both classes are selected for preparing the training samples. Thus, the features from both the original classes are learnt with equal weightage. The downstream model showed an increase in the *True Positive Rate* and a reduction in *Type 2 error*. However, *Type 1 error* increased slightly, subsequently lowering *True Negative Rate*.

TABLE 3.11: Ablation study on the effect of class imbalance on the downstream task in SSLACL.

Model	Accuracy (5%-95% CI)	AUC (5%-95% CI)
without oversampling	76.62 (74.5-78.63)	0.848 (0.828-0.865)
with oversampling	76.72 (74.9-78.70)	0.848 (0.826-0.87)

3.5.4.3 Effect of Number of Parameters in SKID Pretext Model Architecture

TABLE 3.12: Ablation study on the pretext task for Sagittal plane of Magnetic Resonance scans. †The reported pretext validation accuracy is obtained at the epoch with the lowest validation loss. AWGN refers to Additive White Gaussian Noise, which was added during the downstream training phase. 'Y' indicates the addition of noise, while 'N' indicates the absence of noise. $(\mu, \sigma^2) = (0.0, 0.01)$

Id	SKID Model	No. of params	No. of classes	LR decay	Scale	BS	AWGN	Epochs trained	Pretext Validation Accuracy
v1.1	SKID-v1	108.8 M	500	0.95 p.e.	0.25	16	N	50	90.67%†
v1.2	SKID-v1	108.8 M	1000	0.95 p.e.	0.25	16	N	50	85.31%
v2.1	SKID-v2	169.9 M	500	0.95 p.e.	0.25	16	N	50	89.53%
v2.2	SKID-v2	169.9 M	500	0.95 p.e.	0.25	16	Y	50	88.90%
v2.3	SKID-v2	170.4 M	1000	0.95 p.e.	0.25	16	N	50	85.83%
v2.4	SKID-v2	170.4 M	1000	0.95 p.e.	0.25	16	Y	50	85.57%
v3.1	SKID-v3	217.17 M	500	0.95 p.e.	0.25	16	Y	50	94.27%
v3.2	SKID-v3 (Proposed)	217.68 M	1000	0.95 p.e.	0.25	16	Y	50	88.27%

This section aims to compare the performance of the pretext model with different configurations. We obtained these configurations by setting different values of the hyper-parameters like scale, learning rate decay schedule, batch size, etc. This subsection helps to understand how the final model was obtained by repeated experimentation and tuning. SKID-v1 was used as the first model from which we started upgrading the model. In both, SKID-v1 and SKID-v2, a different variant of the *Skip* block was used, where the number of filters in the first convolutional layer is half of the number of channels in its input. The number of output channels in *Skip* and *Dimension Reduction* blocks was also altered to get the different models. The details of the different variants of the architecture are given in Table 3.1. The table shows the number of filters used in building the different architectures that we have experimented upon to obtain the final proposed model (SKID-v3).

Table 3.12 shows the results of the different models under different experimental settings on the pretext task. As we can observe, the results improved as the number of parameters increased. A faster learning rate decay of 0.95 per epoch and a scale factor value of 0.25 proved to be optimal in our experiments. We also used Gaussian noise with a mean of 0.0 and a variance of 0.01 to make our model more robust. Moreover, increasing the batch size helped to achieve convergence faster. The number of classes was increased from 500 to 1000 to train the pretext model to learn sparser features. In the downstream task, the model trained on 1000 classes was used, as the representations learnt by this model are more generalized. The justification for using the model trained with more classes in the pretext task has been justified by an ablation study in Sec. 3.5.4.6. As the number of chosen arrangements increases, the possibility of any convolutional branch learning only some specific patches decreases. This is mainly because the patches become uniformly distributed over the 9 branches, according to the law of large numbers.

However, the primary goal in self-supervised learning is to ensure that the pre-trained models provide a better initialization point for the downstream task. To justify our choice of using SKID-v3 for the downstream experiments, we present the results of the SKID-v1, SKID-v2, and SKID-v3 models in the downstream task of Knee injury classification using only the Sagittal plane MR data. From the illustrative comparative results presented in Fig. 3.13, we can see that the model SKID-v3 performed the best among all the three variations in terms of average AUC score. For this comparison, the models used in the comparison were trained for only 50 epochs and with a batch size of 16 in the pretext task to maintain consistency in the experimental environment. We can infer that increasing the number of parameters of the self-supervised model improves performance on small-scale medical datasets. As all the dataset classes are imbalanced as evident from Fig. 3.8, the AUC scores were used for comparison instead of accuracy.

3.5.4.4 Investigating Label Efficiency in Semi-Supervised Setting in SKID

The quality of the representations learnt by the pretext model can be evaluated by observing its performance in semi-supervised tasks. As the dataset contains only 1130 samples, sampling 1% data from it would only result in 11 samples in the training set. That is why we sample 10% and 50% of the training set to show the efficiency of our model. We follow the same experimental configuration as in the downstream task experiments, where only the classifier part is trained and the parameters in the feature extractor part are not

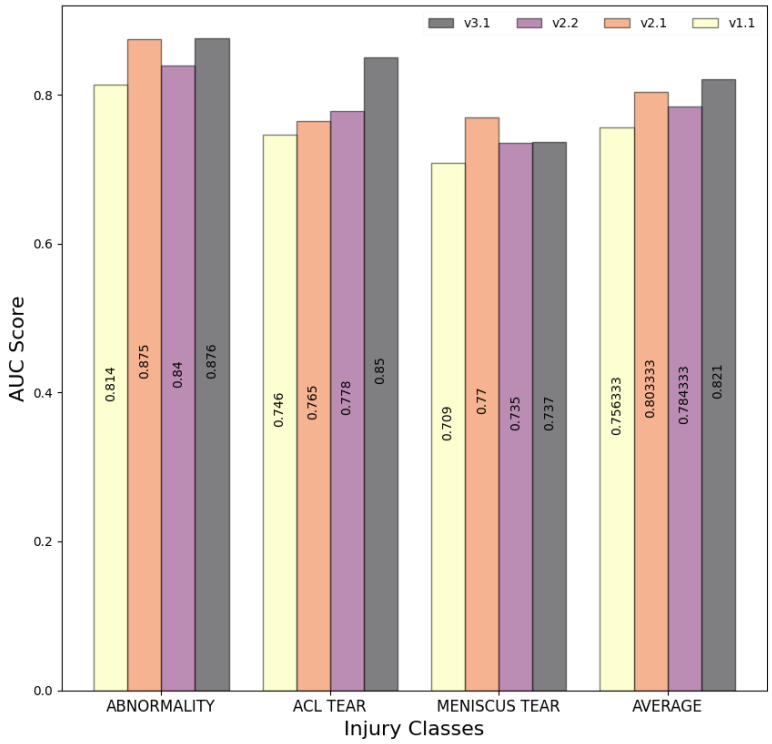


FIGURE 3.13: Comparison of AUC scores in Downstream task of Knee Injury classification using models trained with 500 classes in the pretext stage in the SKID framework.

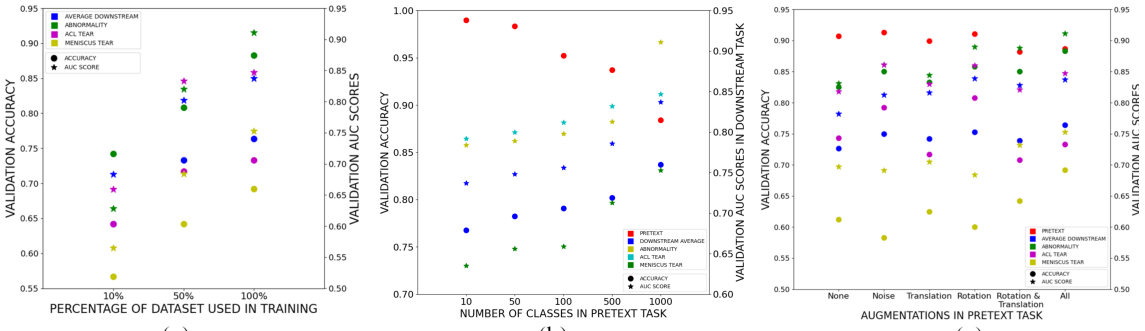


FIGURE 3.14: (a) Validation accuracy and AUC score of downstream models in SKID framework trained on 10%, 50%, and 100% of the dataset, (b) Effect of the number of classes on validation accuracy of pretext and downstream tasks in the SKID framework, (c) Effect of augmentations on the pretext and downstream tasks in the SKID framework. For (a), (b) & (c), circle and asterisk symbols represent validation accuracy and AUC score, respectively. Different colours represent different cases.

updated.

In Fig. 3.14.a, both the validation accuracy and validation AUC scores are plotted for

10%, 50%, and 100% of the training data. It can be observed that, even with training on 10% of the training data, our model performs better than random prediction. Calculating the performance metrics on random predictions, we found the AUC scores for the classes Abnormality, ACL Tear, and Meniscus Tear to be 0.5416 (0.5008-0.5769 5%-95% CI), 0.5955 (0.5697-0.6213 5%-95% CI) and 0.4478 (0.4168-0.4808 5%-95% CI), respectively. Similarly, the accuracy scores for those classes are 0.5312 (0.5025-0.5556 5%-95% CI), 0.5752 (0.5509-0.5981 5%-95% CI), 0.4683 (0.4443-0.4945 5%-95% CI), respectively.

3.5.4.5 ConvLSTM vs 3D CNN: Which Performs Better as a Downstream Classifier in SKID?

For video classification tasks, the use of 3D CNN has been effective. It is common knowledge, that to learn features from temporal data, recurrent neural networks like Long-Short Term Memory (LSTM) serves as an effective tool. Combining convolutional operations in LSTM, we get Convolutional LSTM (ConvLSTM) (Shi et al., 2015). We followed the same experimental configuration for both the models containing ConvLSTM and 3D-CNN (Ji et al., 2013). The comparison results are provided in Table 3.13 and 3.14. It is evident that the ConvLSTM performs better in our experiments and this is the main reason behind preferring ConvLSTM over 3D CNN.

TABLE 3.13: Accuracy comparison between downstream models with ConvLSTM and 3D CNN classifier network in SKID framework.

Method	Accuracy			
	Abnormality	ACL Tear	Meniscus Tear	Average
ConvLSTM	**0.874**	**0.800**	**0.725**	**0.7997**
3D CNN	0.842	0.767	0.708	0.7723

TABLE 3.14: AUC score comparison between downstream models using ConvLSTM and 3D CNN classifier network in SKID framework.

Method	Accuracy			
	Abnormality	ACL Tear	Meniscus Tear	Average
ConvLSTM	**0.904**	**0.893**	**0.810**	**0.869**
3D CNN	0.826	0.830	0.774	0.81

3.5.4.6 Effect of the Number of Classes in the SKID Pretext Task

The generalization capability of the pretext model is directly related to the number of jigsaw arrangements chosen during the training step. In Fig. 3.14.b, we can observe that the AUC scores of the individual classes and also the average accuracy in the downstream task increases as the number of classes increases, even though the validation accuracy in

the pretext task decreases. With the increase in the number of classes, the pretext task becomes more and more difficult, and hence the accuracy of the pretext task decreases. However, it leads to better generalised representation learning and thereby results in better performance in the downstream task.

3.5.4.7 Effect of Augmentations in SKID

In this subsection, we show the effect of different augmentations used in the Pretext task on the performance of the downstream task. Fig. 3.14.c shows the effects of the individual augmentations on the average downstream accuracy and AUC scores, as well as the performance of the individual classes. For the downstream experiments with the different pretext models trained with different augmentations, no augmentation was applied to the input samples. This helps in effectively demonstrating the effect of different augmentations on the representation learnt by the pretext model. The quality of the representations is ultimately reflected in the downstream task performance.

3.6 Conclusion

In this chapter, we discussed two jigsaw puzzle-solving frameworks to learn representations which are context-invariant. Through empirical results, it was observed that the pretext tasks like geometric transformation prediction, rotation prediction, etc. fail to learn proper representations to aid downstream performance and learn low-level representations. It was also found that jigsaw puzzle-solving frameworks are susceptible to this phenomenon. Hence, we proposed a novel architecture which ensures context-invariant representation learning by decoupling the representations from the image patches using a multi-branch convolutional architecture with a channel-wise aggregation strategy. The proposed frameworks outperform the contemporary self-supervised learning as well as supervised learning frameworks.

In the next chapter, we investigate if self-supervised pre-training can enhance the quality of representations learnt in the pretext task. As observed from the results in this chapter, the results of the context-based self-supervised learning frameworks do not outperform the supervised baselines. Hence, to improve the quality of representations and downstream task performance, we resort to contrastive learning principles. In this regard, we propose a binary contrastive learning framework derived from the noise contrastive estimation principle for pre-training on medical images.

Chapter 4

Self-Supervised Contrastive Pre-training on Medical Images

4.1 Introduction

Self-supervised learning in computer vision tasks has evolved from predictive context-based pretext task-solving like geometric transformation prediction (Gidaris et al., 2018; Jing and Tian, 2018), image colorization (Zhang et al., 2016), context prediction (Doersch et al., 2014, 2015), image inpainting (Pathak et al., 2016), temporal order verification (Buckchash and Raman, 2019; Xu et al., 2019; Siar et al., 2020; Misra et al., 2016; Fernando et al., 2017; Liang et al., 2022), frame order prediction (El-Nouby et al., 2019), jigsaw puzzle solving (Noroozi and Favaro, 2016; Kim et al., 2018, 2019; Wei et al., 2019; Ahsan et al., 2019) etc. to optimization-based learning problems. Learning representations from data without human supervision requires mapping the high-dimensional input space to a comparatively low-dimensional space, such that the mapped representations are almost linearly separable. As stated in PIRL (Misra and van der Maaten, 2019), context-based predictive pretext tasks do not help in learning unbiased and context-invariant features. The main reason for such an outcome can be attributed to the limitations of the learning and generalization capability of the model being trained. With manually specified pretext tasks, the model fails to learn features which are not dependent on any specific transformations and often associates features to certain spatial regions as a result of co-dependency between the input data points which may exist in spite of human awareness. This results in a collapse of the representation where the model learns a 'shortcut' to minimize loss function to reach the global minimum by discarding useful information.

Contrastive learning approaches aim at learning representations by maximizing and minimizing the distances between features of data samples belonging to different classes and the same class, respectively. This pushes the feature points of the dissimilar classes further away and pulls the feature points of the same classes closer, thereby creating clusters of features of different classes with maximum separation distance in the feature space. However, in self-supervised contrastive learning (SSCL) algorithms, each input sample is treated as a separate class. In the absence of the ground-truth labels, the optimization

of the contrastive learning objective in SSCL is generally done using a similarity metric on a finite-dimensional feature space. This causes the feature vectors of each sample to be pushed away from each other to the maximum. Besides, the feature vectors of the two samples in a positive pair are pulled closer as well, causing feature vectors of the two different versions of the same sample to be placed as close as possible in the feature space.

There have been several approaches to implement contrastive learning in the self-supervised domain and several algorithms have outperformed supervised baselines in many computer vision tasks. Algorithms such as PIRL (Misra and van der Maaten, 2019), SimCLR (Chen et al., 2020a), MoCo (He et al., 2020) and MoCov2 (Chen et al., 2020c), CPC (van den Oord et al., 2018) and CPCv2 (Hénaff et al., 2020), SwAV (Caron et al., 2020), AIDIM (Bachman et al., 2019), CMC (Tian et al., 2020) all have produced results comparable to supervised algorithms on benchmark natural image datasets. All the above-mentioned algorithms use contrastive loss as the optimizable objective but differ in the way the feature vectors are used. Contrastive learning algorithms require a huge number of negative pairs for effective learning of features. While algorithms like SimCLR (Chen et al., 2020a) or CPC (van den Oord et al., 2018) increase the effective number of negative pairs by algorithmic manipulation, PIRL (Misra and van der Maaten, 2019) and MoCo (He et al., 2020) make use of memory banks to store features vectors. As shown in Wang and Isola (2020) and BYOL (Grill et al., 2020), Euclidean distance is used as a metric for optimization in the feature space. The BYOL (Grill et al., 2020) algorithm however uses only positive pairs for representation learning.

In spite of the tremendous success of contrastive self-supervised learning on natural images, the application of such algorithms on medical data has been limited. Visual medical data differs from natural images in many ways: (1) the features in medical images or videos/scans are present over a small spatial or temporal region, (2) visual medical data is mostly grayscale, and (3) it is hard to obtain annotated data. In self-supervised learning, the representations learned in the pretext task help in providing a better weight initialization for the downstream task, as the dataset used in the pretext and downstream task is the same. This results in the representations being better suited to the downstream task. In the case of transfer learning, the datasets used in the pretext and downstream tasks are different, thus the representations learnt by the network are not adapted to the target dataset used in the downstream task. Furthermore, in transfer learning, if the downstream task dataset is small, then it presents a risk of overfitting and the co-adaptation between the hierarchical features is destroyed, as a small dataset provides fewer samples for the reconstruction of the same (Yosinski et al., 2014).

The primary objective of the framework presented in this chapter is to show that self-supervised pre-training improves performance on the downstream task over supervised baselines in the medical data analysis domain. In the previous chapter, we observed that context-based pretext tasks are susceptible to various issues, and are also unable to outperform the supervised baselines on the downstream tasks. Hence, to improve the quality of representations using the contrastive learning principle, we propose a novel contrastive loss function which does not follow the conventional InfoNCE-based contrastive learning principle. In conventional InfoNCE-based contrastive learning, only the mutual information between the positive pair samples is maximized (van den Oord et al., 2018;

Tian et al., 2020), which is analogous to decreasing the cosine similarity between the samples in all the pairs, as evident from our mathematical analysis. However, the proposed loss function maximizes the cosine similarity between two views of a sample in a positive pair in addition to minimizing the cosine similarity between the samples in all the pairs. To achieve this objective, the contrastive learning task is proposed as a binary classification problem of classifying the pairs as positive or negative. We also present the lower bound analysis of the proposed loss function. Our analysis shows that the lower bound of the proposed loss consists of both alignment and uniformity terms (Wang and Isola, 2020). The above claim is supported by experimental findings presented in this chapter. Furthermore, a squared Euclidean distance-based contrastive loss term similar to the InfoNCE loss is also added to our proposed loss function to improve the representations learnt by our model. Moreover, by pre-training a model with SimCLR (Chen et al., 2020a) on MRNet (Bien et al., 2018) dataset, we observe that self-supervised pre-training improves the representations learnt by the model. We also show that self-supervised pre-training in general with other loss functions also improves performance over the supervised baseline.

The rest of the chapter is organized as follows. Section 4.2 explains the concept behind contrastive learning in brief. In Section 4.3, we discuss the motivation behind the proposed approach. The formulation of the proposed framework and the model architectures of the pretext and downstream experiments are discussed in Section 4.4. Section 4.5 describes the implementation details of the models used in the experiments and also reports comparative results with detailed analyses. Finally, Section 4.6 concludes the paper.

4.2 Preliminaries

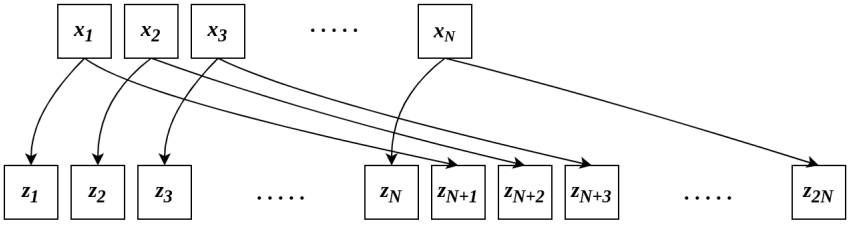


FIGURE 4.1: This figure shows how the feature vectors are obtained from the samples $(x_1, x_2, \ldots, x_N)$ in a batch.

Pair Formation

In Fig. 4.1, we show how the feature vectors are obtained from the samples in a batch and how they are arranged for the final step of calculating the loss. Fig. 4.2 shows how the different pairs are obtained from the feature vectors. We follow the same sampling procedure as in SimCLR (Chen et al., 2020a). Taking a batch size of N, we augment each sample in the batch to obtain two augmented samples from each sample, forming N pairs and $2N$ samples in total. We can form $4N^2$ pairs in total, out of which $T_p = 2N$ are positive pairs and $2N$ are self-pairs, and these $4N$ pairs will not contribute to the contrastive repulsion. Thus, the total number of negative pairs that can be formed is $T_n = 4N^2 - 4N$.

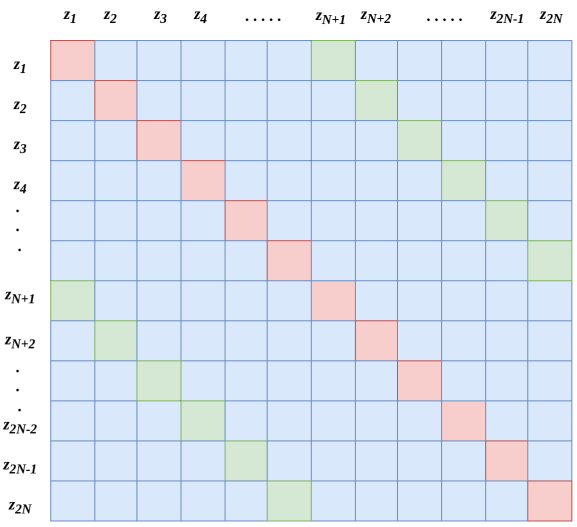


FIGURE 4.2: This figure shows how the pairings are obtained. The red cells indicate self-pairs, green cells indicate positive pairs, i.e., pairings between feature vectors of two augmented versions of the same sample, and blue cells indicate negative pairings, i.e. pairings between feature vectors of different samples.

Contrastive Learning

The contrastive loss function is equivalent to the expected probability of predicting the positive samples correctly. The key point in self-supervised contrastive learning is to treat each sample as a separate class and map feature vectors of samples comprising the positive pair close to each other while repelling the features of samples comprising the negative pair away from each other. This creates an almost linearly separable distribution of features. The separability is further improved by fine-tuning the self-supervised model on the fully or semi-supervised downstream task. The contrastive loss function ($\mathcal{L}_{Cont}$) is generally used in the following form

$$\mathcal{L}_{Cont} = - \mathop{\mathbb{E}}_{\substack{(z^+,z_k)\sim p_+ \\ (z^+,z_i)\sim p_-}} \left[\ln \frac{e^{\langle z^+,z_k\rangle/\tau}}{e^{\langle z^+,z_k\rangle/\tau} + \sum_{\substack{i=1 \\ i\neq k}}^{N} e^{\langle z^+,z_i\rangle/\tau}} \right] \tag{4.1}$$

where $\langle \cdot, \cdot \rangle$ is the cosine similarity between two feature vectors; p_+ and p_- are the distribution of positive pairs and negative pairs on $\mathbb{R}^n \times \mathbb{R}^n$, respectively, and τ is the temperature parameter. N is the total number of pairs in a single batch of training. Here z^+ is assumed to be the anchor sample.

Let us assume, $\mathcal{X}_+$ and $\mathcal{X}_-$ as the sets of all positive and negative pairs, respectively. In a training batch, the elements belonging to $\mathcal{X}_+$ and $\mathcal{X}_-$ are drawn from p_+ and p_-, respectively. Thus, we can say that the positive and negative pairs are drawn from the sets $\mathcal{X}_+$ and $\mathcal{X}_-$ instead of the distributions p_+ and p_-. We express $\mathcal{X}_{pair} = \mathcal{X}_+ \cup \mathcal{X}_-$ as the union of all pairs (both positive and negative) and the corresponding distribution as p_{pair}. This allows us to rewrite the Eqn 4.1 in the form given below

$$\mathcal{L}_{Cont} = -\frac{1}{N} \sum_{\substack{(z^+,z_k)\in\mathcal{X}_+ \\ (z^+,z_i)\in\mathcal{X}_-}} \ln \frac{e^{\frac{\langle z^+,z_k\rangle}{\tau}}}{e^{\frac{\langle z^+,z_k\rangle}{\tau}} + \sum_{\substack{(z^+,z_i)\in\mathcal{X}_- \\ i\neq k}} e^{\frac{\langle z^+,z_i\rangle}{\tau}}} \tag{4.2}$$

Since,

$$e^{\frac{\langle z^+,z_k\rangle}{\tau}} + \sum_{\substack{(z^+,z_i)\in\mathcal{X}_- \\ i\neq k}} e^{\frac{\langle z^+,z_i\rangle}{\tau}} = \sum_{(z^+,z_i)\in\mathcal{X}_+\cup X_-} e^{\frac{\langle z^+,z_i\rangle}{\tau}} = \sum_{(z^+,z_i)\in\mathcal{X}_{pair}} e^{\frac{\langle z^+,z_i\rangle}{\tau}} \tag{4.3}$$

Putting, the above derivation in Eqn. 4.2, we get

$$\mathcal{L}_{Cont} = -\frac{1}{N} \sum_{\substack{(z^+,z_k)\in\mathcal{X}_+ \\ (z^+,z_i)\in\mathcal{X}_{pair}}} \ln \frac{e^{\frac{\langle z^+,z_k\rangle}{\tau}}}{\sum_{(z^+,z_i)\in\mathcal{X}_{pair}} e^{\frac{\langle z^+,z_i\rangle}{\tau}}} \tag{4.4}$$

This can also be written as,

$$\mathcal{L}_{Cont} = -\frac{1}{N} \sum_{i}^{N} \ln \frac{e^{\frac{\langle z_i,z_i^+\rangle}{\tau}}}{\sum_{(z_k,z_l)\in\mathcal{X}_{pair}} e^{\frac{\langle z_k,z_l\rangle}{\tau}}} \tag{4.5}$$

Lower Bound Analysis of InfoNCE loss

In this subsection, we will analyse the asymptotic behaviour of the InfoNCE loss function. We found that the InfoNCE loss can be reduced such that it is lower bound tightly by the difference of the From Eqn. 4.5 we get,

$$\begin{aligned}
\mathcal{L}_{Cont} &= -\frac{1}{N} \sum_{i}^{N} \ln \frac{e^{\frac{<z_i,z_i^+>}{\tau}}}{\sum_{(z_k,z_l)\in\mathcal{X}_{pair}} e^{\frac{<z_k,z_l>}{\tau}}} \\
&= -\frac{1}{N} \sum_{i}^{N} \frac{<z_i,z_i^+>}{\tau} + \frac{1}{N} \sum_{i}^{N} \ln \left(\sum_{(z_k,z_l)\in\mathcal{X}_{pair}} e^{\frac{<z_k,z_l>}{\tau}} \right) \\
&= -\frac{1}{N} \sum_{i}^{N} \left[\frac{<z_i,z_i^+>}{\tau} - \ln \left(\sum_{(z_k,z_l)\in\mathcal{X}_{pair}} e^{\frac{<z_k,z_l>}{\tau}} \right) \right] \\
&\geq -\frac{1}{N} \sum_{i}^{N} \left[\frac{<z_i,z_i^+>}{\tau} - \sum_{(z_k,z_l)\in\mathcal{X}_{pair}} \ln(e^{\frac{<z_k,z_l>}{\tau}}) \right] \text{ [Using Jensen's Inequality]} \\
&\geq -\frac{1}{N} \sum_{i}^{N} \left[\frac{<z_i,z_i^+>}{\tau} - \sum_{(z_k,z_l)\in\mathcal{X}_{pair}} \frac{<z_k,z_l>}{\tau} \right] \\
&\geq -\frac{1}{N} \sum_{i}^{N} \left[\frac{<z_i,z_i^+>}{\tau} - \sum_{(z_k,z_l)\in\mathcal{X}_+\cup\mathcal{X}_-} \frac{<z_k,z_l>}{\tau} \right]
\end{aligned} \tag{4.6}$$

$$
\begin{aligned}
&\geq -\frac{1}{N}\sum_{i}^{N}\frac{< z_i, z_i^+ >}{\tau} + \frac{1}{N}\sum_{i}^{N}\sum_{(z_k,z_l)\in\mathcal{X}_+\cup\mathcal{X}_-}\frac{< z_k, z_l >}{\tau} \\
&\geq -\frac{1}{N}\sum_{i}^{N}\frac{< z_i, z_i^+ >}{\tau} + \frac{1}{N}\sum_{i}^{N}\left[\sum_{(z_m,z_n)\in\mathcal{X}_+}\frac{< z_m, z_n >}{\tau} + \sum_{(z_k,z_l)\in\mathcal{X}_-}\frac{< z_k, z_l >}{\tau}\right] \\
&\geq \frac{N-1}{N}\sum_{i}^{N}\frac{< z_i, z_i^+ >}{\tau} + \frac{1}{N}\sum_{i}^{N}\sum_{(z_k,z_l)\in\mathcal{X}_-}\frac{< z_k, z_l >}{\tau} \\
&\geq \sum_{i}^{N}\frac{< z_i, z_i^+ >}{\tau} + \mathbb{E}\left[\sum_{(z_k,z_l)\in\mathcal{X}_-}\frac{< z_k, z_l >}{\tau}\right] \text{ [for large } N\text{]} \\
&\approx \sum_{i}^{N}\frac{< z_i, z_i^+ >}{\tau} + \sum_{(z_k,z_l)\in\mathcal{X}_-}\frac{< z_k, z_l >}{\tau} \text{ [for large } N\text{]} \\
&\approx \sum_{(z_i,z_i^+)\in\mathcal{X}_+}\frac{< z_i, z_i^+ >}{\tau} + \sum_{(z_k,z_l)\in\mathcal{X}_-}\frac{< z_k, z_l >}{\tau} \text{ [for large } N\text{]} \\
&\approx \sum_{(z_i,z_j)\in\mathcal{X}_{pair}}\frac{< z_i, z_j >}{\tau} \text{ [for large } N\text{]}
\end{aligned}
$$

The contrastive loss can be expanded as in the above derivation Eqn. 4.2. It can be observed that optimizing a contrastive loss function in a self-supervised scenario is asymptotically the same as minimizing the cosine similarity between every sample where each sample is treated as a separate class. The lower bound of the InfoNCE loss is the same as the uniformity metric term in contrastive learning (Wang and Isola, 2020).

Unnormalized Statistical Models

As described in Gutmann and Hyvärinen (2012), the basic estimation problem is formulated as follows. Assume a sample of a random vector $x \in \mathbb{R}^n$ is observed which follows an unknown probability density function (pdf) $p_d(\cdot)$. The data pdf $p_d(\cdot)$ is modeled by a parameterized family of functions $\{p_m(.;\theta)\}_\theta$ where θ is a vector of parameters. We assume that $p_d(\cdot)$ belongs to this family. In other words, $p_d(\cdot) = p_m(\cdot;\theta^*)$ for some parameter θ^*. To estimate θ from the observed sample, we optimize some objective functions.

Any solution $\hat{\theta}$ to the estimation problem yields a properly normalized density $p_m(\cdot;\hat{\theta})$ with

$$\int p_m(x;\hat{\theta})dx = 1 \tag{4.7}$$

This is essentially a constraint in the optimization problem. In principle, the constraint can always be fulfilled by redefining the pdf as

$$p_m(\cdot;\theta) = \frac{p_m^0(\cdot;\theta)}{Z_\theta}$$
$$\text{where } Z_\theta = \int p_m^0(dx;\theta)dx \tag{4.8}$$

and $p_m^0(dx;\theta)$ specify the functional form of the pdf and do not need to integrate into one.

4.3 Motivation

The primary motivation for the proposed loss is to classify the pairs as positive or negative. This modifies the problem into a binary classification problem. In this subsection, we justify the formulation of the proposed loss from the noise contrastive estimation principle (Gutmann and Hyvärinen, 2012).

Let us denote the set of positive and negative pairs as $\mathcal{X}_+$ and $\mathcal{X}_-$, respectively. A pair (z_i, z_j) is assigned a binary class label k_{ij}: $k_{ij} = 1$ if $(z_i, z_j) \in \mathcal{X}_+$ and $k_{ij} = 0$ if $(z_i, z_j) \in \mathcal{X}_-$.

In self-supervised learning, the distribution of the data, as well as the distribution of the samples are also unknown. While this objective is similar to the Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2012), the formulation of our first loss follows NCE in some aspects. We will discuss the same in the following paragraphs. First, we will discuss the basic notions of NCE, and then discuss the steps to obtain the proposed framework.

Deriving Proposed Loss using the notion of Noise Contrastive Estimation

Let us define the class-conditional probability densities as follows,

$$\begin{aligned} p((z_i, z_j)|k=1) &= p_m((z_i, z_j);\theta) \\ p((z_i, z_j)|k=0) &= p_n((z_i, z_j)) \end{aligned} \tag{4.9}$$

where $p_m(\cdot;\theta)$ and $p_n(\cdot)$ are the normalized parameterized data distribution and noise distribution, respectively, as already discussed in Sec. 4.2.

When the samples in a batch with size N are paired as shown in Sec. 4.2, then we get $4N^2 - 4N$ negative pairs, and $2N$ positive pairs. We discard the $2N$ number of self-pairs. Therefore, the prior probabilities should be as follows,

$$\begin{aligned} P(k=1) &= \frac{2N}{4N^2 - 2N} = \frac{1}{2N-1} \\ P(k=0) &= \frac{4N^2 - 4N}{4N^2 - 2N} = \frac{2N-2}{2N-1} \end{aligned} \tag{4.10}$$

Following Gutmann and Hyvärinen (2012), we observe that the value of the imbalance factor $\nu = 2N - 2$. Hence, the posterior probabilities for the classes can be defined as follows,

$$\begin{aligned} P(k=1|(z_i,z_j);\theta) &= \frac{p_m((z_i,z_j);\theta)}{p_m((z_i,z_j);\theta) + \nu p_n((z_i,z_j))} \\ P(k=0|(z_i,z_j);\theta) &= \frac{\nu p_n((z_i,z_j))}{p_m((z_i,z_j);\theta) + \nu p_n((z_i,z_j))} \end{aligned} \tag{4.11}$$

Denoting $P(k=1|(z_i,z_j);\theta)$ by $\phi((z_i,z_j);\theta)$, the loss function can be defined as follows

$$\begin{aligned} \mathcal{L} &= -\frac{1}{T_p+T_n} \sum_{(z_i,z_j)\in\mathcal{X}_+\bigcup\mathcal{X}_-} [k_{ij}\ln P(k_{ij}=1|(z_i,z_j);\theta) + (1-k_{ij})\ln P(k_{ij}=0|(z_i,z_j);\theta)] \\ &= -\frac{1}{T_p+T_n}\left[\sum_{(z_i,z_j)\in\mathcal{X}_+} \ln\phi((z_i,z_j);\theta) - \sum_{(z_k,z_l)\in\mathcal{X}_-} \ln(1-\phi((z_k,z_l);\theta)\right] \end{aligned} \tag{4.12}$$

where $\phi((z_i,z_j);\theta)$ can be written as

$$\phi((z_i,z_j);\theta) = \frac{1}{1+\nu\exp(-\mathcal{G}_{NC}((z_i,z_j);\theta)} \tag{4.13}$$

where $\mathcal{G}_{NC}((z_i,z_j);\theta) = \ln p_m((z_i,z_j);\theta) - \ln p_n((z_i,z_j)) = \ln\frac{p_m((z_i,z_j);\theta)}{p_n((z_i,z_j))}$, which is the log-odds of the pair (z_i,z_j) belonging to the normalized parameterized data distribution $p_m(\cdot|\theta)$ or the noise distribution $p_n(\cdot)$.

For the proposed loss function we choose $\mathcal{G}_{NC}((z_i,z_j);\theta)$ such that it gives the logit value for the non-parametric sigmoid classifier (logistic regression), that is, $\mathcal{G}_{NC}((z_i,z_j);\theta) = \frac{1}{\tau}\left(z_i\cdot z_j^T\right) = \frac{c_{ij}}{\tau}$. c_{ij} denotes the cosine similarity between z_i and z_j and is the SSL equivalent to logit values in non-parametric logistic regression. τ is the temperature hyper-parameter.

Cosine Similarity instead of log-odds of probability: The value of the log-odds of the probability of a pair belonging to the normalized parametrized data distribution $p_m(\cdot;\theta)$ and the noise distribution $p_n(\cdot)$, that is, $\ln\frac{p_m((z_i,z_j);\theta)}{p_n((z_i,z_j))} = \mathcal{G}_{NC}$ theoretically can range from $-\infty$ to $+\infty$. Furthermore, the value of $\mathcal{G}_{NC}$ increases as the probability of the (z_i,z_j) belonging to $p_m(\cdot;\theta)$ increases, and vice versa. Similarly, if the pair (z_i,z_j) is sampled from $p_m(\cdot;\theta)$, the cosine similarity between z_i and z_j will be high, and if it is sampled from the noise distribution $p_n(\cdot)$ then the cosine similarity will be low. As cosine similarity is bounded between -1 to $+1$, we can scale the range by using a temperature hyperparameter τ. Taking into account the above discussion, we can clearly state without any loss of generality, that $\ln\frac{p_m((z_i,z_j);\theta)}{p_n((z_i,z_j))} = \frac{1}{\tau}\left(z_i\cdot z_j^T\right)$.

Based on the motivation discussed in this section, in the next section, we present the proposed loss function, its lower bound analysis, and the model architectures used in the pretext and downstream tasks.

4.4 Proposed Framework

In this section, we present the proposed binary contrastive framework. We then present the asymptotic analysis of the proposed loss and show that it is capable of optimizing the cosine similarity between the samples in the positive pairs in addition to minimizing the cosine similarity between the samples in all pairs. In addition to that, we present another loss which is a combination of the proposed loss with the InfoNCE loss to further improve the quality of representations. This is followed by a discussion of the model architectures used in the pretext and downstream tasks.

4.4.1 Binary Contrastive Loss

In this subsection, we present the proposed loss formulation and show the lower bound analysis of the same. We also present the formulation of the loss obtained by combining the proposed loss and the InfoNCE loss to empirically observe if the combined effect of both the loss can improve performance further.

Considering N samples in each batch, transforming each sample in two ways gives us $2N$ samples or N pairs. With $2N$ samples we can form $4N^2$ pairs with each pair occurring twice. Taking the same notations as used in Eqn. 4.1, the loss function proposed in this work is presented in Eqn. 4.14,

$$\begin{aligned}\mathcal{L}_{Proposed} &= -\mathop{\mathbb{E}}_{(x_i,x_j)\sim p_{pair}}\left[y_{ij}\ln\left(\frac{1}{1+e^{\frac{-<x_i,x_j>}{\tau}}}\right)+(1-y_{ij})\ln\left(1-\frac{1}{1+e^{\frac{-<x_i,x_j>}{\tau}}}\right)\right]\\ &= -\frac{1}{4N^2}\left[\sum_{(x_i,x_j)\in\mathcal{X}_+}\ln\left(\frac{1}{1+e^{\frac{-<x_i,x_j>}{\tau}}}\right)+\sum_{(x_k,x_l)\in\mathcal{X}_-}\ln\left(1-\frac{1}{1+e^{\frac{-<x_k,x_l>}{\tau}}}\right)\right]\end{aligned} \tag{4.14}$$

where

$$y_{ij}=\begin{cases}1,\text{when}(x_i,x_j)\in\mathcal{X}_+\\0,\text{when}(x_i,x_j)\in\mathcal{X}_-\end{cases}$$

Since, in a contrastive learning scenario there exists only two types of pair: positive and negative, we intend to model contrastive learning as a binary classification problem. The loss function is the total expectation of the log of the probability of correctly predicting a pair, either positive or negative. We can infer that minimizing the second term means that the model essentially learns to map the features of the samples in a negative pair into points in the feature space where the dependency between them is minimized.

4.4.1.1 Lower Bound Analysis of the Proposed Loss

In this section, we mathematically analyse the lower bound of the proposed binary contrastive loss. We will observe that the lower bound of the proposed loss is a difference between the alignment and uniformity terms. The mathematical analysis is shown below.

From Eqn. 4.14 we get,

$$
\begin{aligned}
\mathcal{L}_{Proposed} &= -\frac{1}{4N^2}\left[\sum_{(x_i,x_j)\in\mathcal{X}_+}\ln\left(\frac{1}{1+e^{-<x_i,x_j>/\tau}}\right)+\sum_{(x_k,x_l)\in\mathcal{X}_-}\ln\left(1-\frac{1}{1+e^{-<x_k,x_l>/\tau}}\right)\right] \\
&= -\frac{1}{4N^2}\left[\sum_{(x_i,x_j)\in\mathcal{X}_+}\ln\left(\frac{1}{1+e^{-<x_i,x_j>/\tau}}\right)+\sum_{(x_k,x_l)\in\mathcal{X}_-}\ln\left(1-\frac{1}{1+e^{-<x_k,x_l>/\tau}}\right)\right] \\
&= -\frac{1}{4N^2}\left[\sum_{(x_i,x_j)\in\mathcal{X}_+}\ln\left(\frac{e^{<x_i,x_j>/\tau}}{1+e^{<x_i,x_j>/\tau}}\right)+\sum_{(x_k,x_l)\in\mathcal{X}_-}\ln\left(\frac{1}{1+e^{<x_k,x_l>/\tau}}\right)\right] \\
&= -\frac{1}{4N^2}\left[\sum_{(x_i,x_j)\in\mathcal{X}_+}\left(\frac{<x_i,x_j>}{\tau}-\ln\left(1+e^{<x_i,x_j>/\tau}\right)\right)-\sum_{(x_k,x_l)\in\mathcal{X}_-}\ln\left(1+e^{<x_k,x_l>/\tau}\right)\right] \\
&\quad [\text{Putting } <x_i,x_j> \text{ (cosine similarity) as } c_{ij}] \\
&= -\frac{1}{4N^2}\left[\sum_{(x_i,x_j)\in\mathcal{X}_+}\frac{c_{ij}}{\tau}-\sum_{(x_i,x_j)\in\mathcal{X}_+}\ln\left(1+e^{c_{ij}/\tau}\right)-\sum_{(x_k,x_l)\in\mathcal{X}_-}\ln\left(1+e^{c_{kl}/\tau}\right)\right] \\
&\quad [\text{Putting } \ln(1+x)=\ln(2)+\frac{x}{2}+\mathcal{O}(x^2) \text{ and ignoring the higher order terms}] \\
&\geq -\frac{1}{4N^2}\left[\sum_{(x_i,x_j)\in\mathcal{X}_+}\frac{c_{ij}}{\tau}-\sum_{(x_k,x_l)\in\mathcal{X}_{pair}}\left(\ln(2)+\frac{c_{kl}}{2\tau}\right)\right] \\
&\geq -\frac{1}{4N^2}\left[\sum_{(x_i,x_j)\in\mathcal{X}_+}\frac{c_{ij}}{\tau}-\sum_{(x_k,x_l)\in\mathcal{X}_{pair}}\frac{c_{kl}}{2\tau}-(4N^2-4N)\ln(2)\right] \\
&\geq -\frac{1}{4N^2}\sum_{(x_i,x_j)\in\mathcal{X}_+}\frac{c_{ij}}{\tau}+\frac{1}{4N^2}\sum_{(x_k,x_l)\in\mathcal{X}_{pair}}\frac{c_{kl}}{2\tau}+C \quad [\text{For large } N]
\end{aligned}
\tag{4.15}
$$

where C is an additive constant.

As shown above, the lower bound of the proposed loss function can be expressed as follows

$$
\mathcal{L}_{Proposed} \geq -\frac{1}{4N^2}\left[\sum_{(x_i,x_j)\in\mathcal{X}_+}\frac{<x_i,x_j>}{\tau}-\sum_{(x_k,x_l)\in\mathcal{X}_{pair}}\frac{<x_k,x_l>}{2\tau}\right]+C \tag{4.16}
$$

The order of magnitude of first and second terms in the above expression, in terms of N is $\mathcal{O}(\frac{1}{N})$ and $\mathcal{O}(1)$, respectively. The proposed loss not only minimizes the cosine similarity

between the samples in all the pairs but also increases the cosine similarity between the samples in the positive pairs. The first term in the lower bound is the same as the alignment metric and the second term is the uniformity metric (Wang and Isola, 2020). Thus, the proposed binary contrastive loss influences both the alignment and uniformity.

4.4.1.2 Combining with InfoNCE Loss

In this chapter, we also look to study the contribution of both InfoNCE and the proposed framework and also compare with the contemporary state-of-the-art by using 3 types of loss functions for different experiments. The first one is the loss function that we propose ($\mathcal{L}_{Proposed}$). The second loss function that used in this chapter is the linear combination of the squared Euclidean distance-based contrastive loss function $\mathcal{L}_{Contr.}$ shown in Eqn. 4.17 and $\mathcal{L}_{Proposed}$. Combining the two, we obtain $\mathcal{L}_{Combo}$ (Eqn. 4.18) used in our experiments. The last one is the loss function used in SimCLR (Chen et al., 2020a), $\mathcal{L}_{SimCLR}$. We also use $\mathcal{L}_{SimCLR}$ for comparison with our proposed loss function in the downstream task.

$$\mathcal{L}_{Contr} = - \mathop{\mathbb{E}}_{\substack{(z^+,z_k)\sim p_+ \\ (z^+,z_i)\sim p_-}} \left[\ln \frac{e^{-\frac{\|z^+ - z_k\|_2^2}{\tau_c}}}{e^{-\frac{\|z^+ - z_k\|_2^2}{\tau_c}} + \sum_{\substack{i=1 \\ i\neq k}}^{N} e^{-\frac{\|z^+ - z_i\|_2^2}{\tau_c}}} \right] \tag{4.17}$$

The expression for $\mathcal{L}_{Combo}$ used in our experiments is as given below:

$$\begin{aligned} \mathcal{L}_{Combo} =& \mathcal{L}_{Proposed} + \mathcal{L}_{Contr.} \\ =& - \mathop{\mathbb{E}}_{(x_i,x_j)\sim p_+} \left[\ln \left(\frac{1}{1 + e^{-<x_i,x_j>/\tau}} \right) \right] - \mathop{\mathbb{E}}_{(x_k,x_l)\sim p_-} \left[\ln \left(1 - \frac{1}{1 + e^{-<x_k,x_l>/\tau}} \right) \right] \\ & - \mathop{\mathbb{E}}_{\substack{(z^+,z_k)\sim p_+ \\ (z^+,z_i)\sim p_-}} \left[\ln \frac{e^{-\frac{\|z^+ - z_k\|_2^2}{\tau_c}}}{e^{-\frac{\|z^+ - z_k\|_2^2}{\tau_c}} + \sum_{\substack{i=1 \\ i\neq k}}^{N} e^{-\frac{\|z^+ - z_i\|_2^2}{\tau_c}}} \right] \end{aligned} \tag{4.18}$$

4.4.2 Model Architecture

Self-supervised learning consists of two parts: Pretext and Downstream. Thus, we use two different architectures, depending on the stage of our task. In the pretext task, we use the architecture shown in Fig. 4.3. It consists of an Encoder and a Projector. For the encoder, we use a ResNet (He et al., 2016) model. The output from the ResNet model is taken from the average pooling layer. The dimensions of the output from the ResNet encoder are $N_S \times 2048 \times 1 \times 1$, where N_S is the number of slices in a single batch. The projector is a 2-layered multi-layer perceptron. The input and output dimensions of the first linear layer are 2048 and are followed by a ReLU (He et al., 2015) layer. The output dimension of the second linear layer is 512 or 1024, depending on the model.

For the downstream task, we take the Encoder only and discard the Projector. To the output of the Encoder, we add a custom layer which computes the maximum of the features obtained from the Encoder, over the slices. As mentioned in Sec. 4.5.2, we use only 16 slices for the downstream task from each sample. Thus, the dimensions of the output from the Encoder is $(N \times 16) \times 2048 \times 1 \times 1$, where N is the number of samples in each batch. We take the maximum of the features over 16 slices to obtain an output of dimensions $N \times 2048$. A linear layer of dimensions 2048×1 is added to its output for obtaining the final predictions.

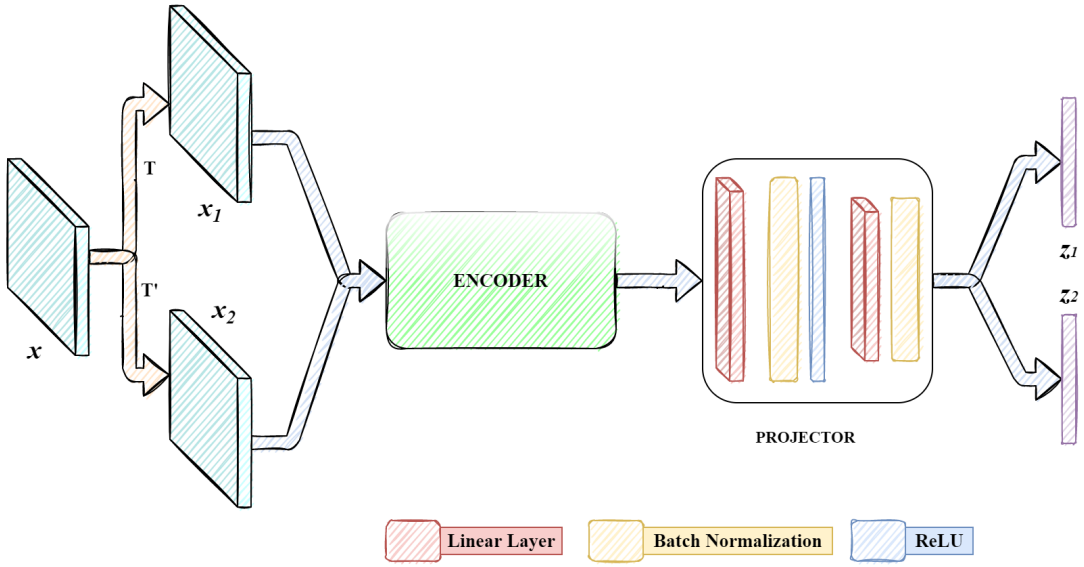


FIGURE 4.3: Proposed pretext model architecture.

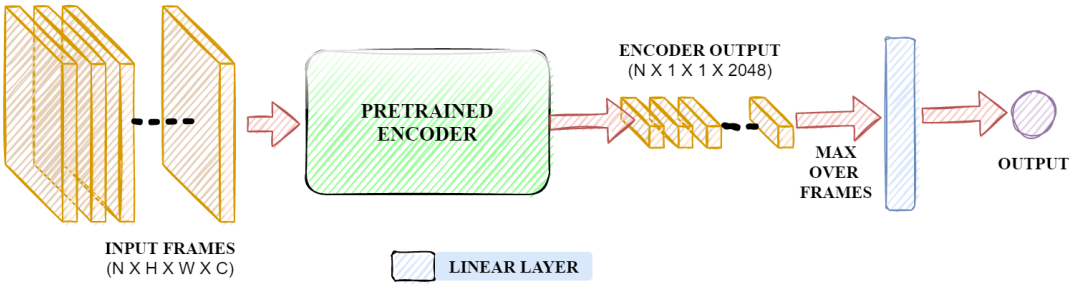


FIGURE 4.4: Architecture of the model used in the downstream tasks.

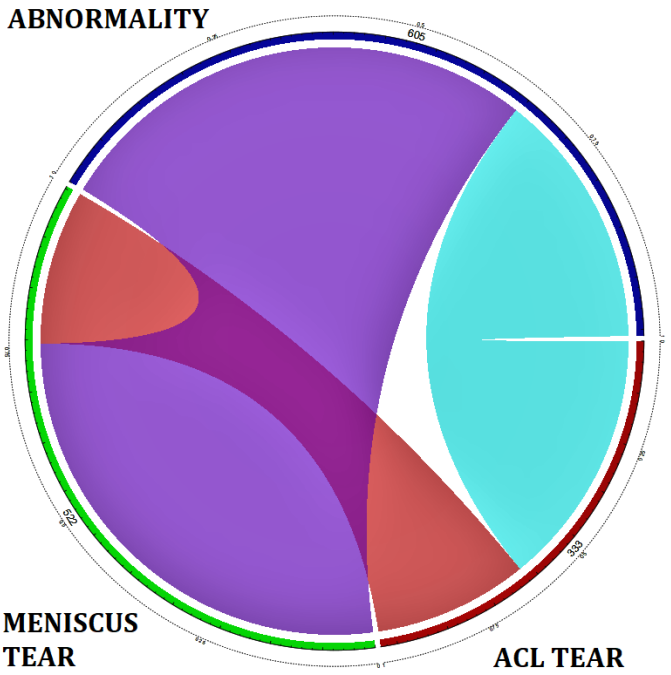


FIGURE 4.5: Concurrence plot of MRNet dataset showing label concurrence in the imbalanced multilabel dataset MRNet (Bien et al., 2018).

4.5 Experimental Details, Results and Analysis

4.5.1 Dataset

For this work, we use the MRNet (Bien et al., 2018) dataset. This dataset consists of magnetic resonance scans of the human knee. The training set consists of 1130 scans each for the three planes: Sagittal, Coronal and Axial. The validation set consists of 120 scans for each of the three planes. This dataset is also a multilabel dataset and the individual labels are highly imbalanced, which makes the work more challenging. In Table 4.1, the SCUMBLE (Charte et al., 2019) scores for the MRNet (Bien et al., 2018) dataset is given to understand its data distribution. The SCUMBLE (Charte et al., 2019) metric aims to quantify the imbalance variance among the labels present in each data sample. Hence, we use it to understand the nature of the MRNet (Bien et al., 2018) dataset. The Fig. 4.5, shows the interaction of different labels with each other using a concurrence plot.

TABLE 4.1: SCUMBLE scores of MRNet dataset

SCUMBLE				
Mean	CV	Abnormality	ACL Tear	Meniscus Tear
0.05314746	1.344904	0.183	0.1048	0.06578

4.5.2 Implementation Details

In this subsection, we discuss the implementation details of both pretext and downstream tasks of both the proposed frameworks.

4.5.2.1 Pretext Implementation Details

The pretext models were optimized using LARS (You et al., 2017) optimizer with an initial learning rate of 0.1. The learning rate was decayed following a Cosine decay schedule without restarts and also without any warmup. Based on our experiments, the temperature parameter τ in $\mathcal{L}_{Proposed}$ is set to 0.25. For the temperature parameter τ_c in $\mathcal{L}_{Contr.}$ in Eqn. 4.17, we set the value to 0.33. We conducted the experiments using two variants of ResNet (He et al., 2016), namely, ResNet50 and ResNet101. The output dimension of the projector was varied between 512 and 1024. The primary reason behind using a higher-dimension feature vector is to increase the sparsity of the mapped data points, thereby increasing the separability of the samples. In Table 4.2, the different configurations of the models used in our pretext experiments are given.

For the pretext training process, we followed a procedure similar to SimCLR (Chen et al., 2020a). We sampled a slice $\mathcal{F}$ randomly from each MR scan and then augmented the sampled slice in two different ways $\mathcal{T}_1$ and $\mathcal{T}_2$ to give a positive pair $(\mathcal{T}_1(\mathcal{F}), \mathcal{T}_2(\mathcal{F}))$. Thus, for a batch size of 64, 128 samples are obtained. This gives us $128^2 = 16384$ pairs. Out of these 16384 pairs, we discard $2 \times 128 = 256$ pairs, because those are pairs of samples with each other. After calculating the cosine similarity of each pair and putting the cosine similarity values in Eqn. 4.14, we calculate the binary cross-entropy loss of the positive pair and negative pair being correctly classified. Ideally, the cosine similarity for a positive pair should be 1 and that for the negative pair should be -1 and the binary classifier should predict 1 and 0 for positive and negative pairs, respectively. The pretext training was run for only 100 epochs on an NVIDIA P100 GPU on Google Colab. The time required for training a ResNet50 model on the MRNet dataset for 100 epochs is approximately 1.5 hours. In the pre-training phase, we trained a single model for each plane: Sagittal, Coronal and Axial. The augmentations applied are mentioned in Table 4.3.

TABLE 4.2: Details of different models used in the pretext experiments

Model No.	Base Encoder	Feature Dimension	Batch Size	Loss Function
1 (Proposed)	ResNet50	512	64	$\mathcal{L}_{Proposed}$
2 (Proposed)	ResNet50	1024	64	$\mathcal{L}_{Proposed}$
3 (Combo)	ResNet50	512	64	$\mathcal{L}_{Combo}$
4 (Combo)	ResNet50	1024	64	$\mathcal{L}_{Combo}$
5 (Combo)	ResNet101	1024	32	$\mathcal{L}_{Combo}$
6 (SimCLR)	ResNet50	512	64	$\mathcal{L}_{SimCLR}$

In the pretext experiments, we apply augmentations to each slice to obtain two different views of the slice. These two different views of a slice form a positive pair. The augmentations we used are given in Table 4.3.

TABLE 4.3: Augmentations used in the pretext experiments. H and W are the height and width of a slice of an MR scan.

Augmentation	Value
Random Cropping with Resizing	window = (224, 224) scale = (0.08, 1.0)
Gaussian Blur	kernel = (5,5) sigma=(0.1, 2.0)
Random Rotation	range = $(-30.0°, +30.0°)$
CutOut (Devries and Taylor, 2017)	number = randint(1,4) dimension = $\lfloor min(H, W)/4 \rfloor$
Color Distortion	Brightness = 0.2 Contrast = 0.2 Saturation = 0.2 Hue = 0.0
Gaussian Noise	mean = 0.0 standard deviation = 0.1

4.5.2.2 Downstream Implementation Details

In the downstream task, for handling the multi-label nature of the data, we transformed the problem into separate binary classification tasks (Godbole and Sarawagi, 2004; Tsoumakas and Katakis, 2007). We trained 9 models, three for each of the three planes: Sagittal, Coronal and Axial. Among the three models for each plane, one model is trained for each of the three labels: Abnormality, ACL Tear and Meniscus Tear. As the number of samples is already low in the training set, we chose to fine-tune the whole model for the downstream task by optimizing the binary cross-entropy loss using Adam (Kingma and Ba, 2015) optimizer with an initial learning rate of 10^{-5}, decayed exponentially at the rate of 0.9 per epoch for 10 epochs only.

In the downstream experiments, we do not use any augmentations except random cropping with resizing to 224×224. Because of GPU memory limitation, we sampled only 16 slices randomly from an MR scan and also limited the batch size to 4. Training time for each model is approximately 20 minutes. The total training time for both the pretext and the downstream models is approximately $1.5 \times 3 + \frac{20}{60} \times 9 = 7.5$ hours.

To obtain the final output we did an ensemble of the nine models using a weighted average of the prediction probabilities of the three models. For the ensemble, we used weighted soft voting (Zhou, 2012) approach similar to the ensemble strategy in Sec. 3.4.2.3. The performance of the 3 models would not be the same for any of the classes. Thus, it is essential to give more weight to the stronger classifier for each class. The output prediction of the ensemble of the 3 models for the j^{th} class is given by

$$\hat{y} = \sum_{i=1}^{N_C} w_i^j f_i^j(x) > 0.5 \tag{4.19}$$

where w_i^j is the weight assigned to the classifier output f_i^j for the j^{th} class by the i^{th} classifier, and N_C is the total number of classifiers. The weights are non-negative and are constrained by $\sum_{i=1}^{N_C} w_i^j = 1$. The weights are calculated as

$$w_i^j = \ln(\frac{acc_i^j}{1 - acc_i^j}) \tag{4.20}$$

where acc_i^j is the prediction accuracy for the j^{th} class by the i^{th} classifier.

Since we select only 16 slices randomly from an MR scan for each sample, we follow a Monte Carlo method to infer the final predictions on the validation set similar to the one done in Sec. 3.4.2.3. The predictions $f_i^j(x)$ are calculated by taking the average over 8 different samples of 16 slices, sampled randomly from an MR scan similar to Eqn. 3.3. The reason for doing this is to decrease the uncertainty associated with the prediction probability by feeding the model with more information, at the same time keeping each sample temporally sparse.

4.5.3 Comparative Results and Analysis

In this section, we present the results of the different models mentioned in the previous subsection in the downstream task of Knee MR diagnosis or injury classification on the MRNet dataset.

4.5.3.1 Results of Model Pre-trained with $\mathcal{L}_{Proposed}$

In Table 4.4 and 4.5, we present the results of applying the proposed loss function $\mathcal{L}_{Proposed}$ in the pre-training phase. In Table 4.4, we show the accuracy of the ensemble of fine-tuned pre-trained ResNet (He et al., 2016) models on the downstream task, that is multilabel classification on the MRNet dataset. In Table 4.5, we present the AUC scores for the same models.

TABLE 4.4: Accuracy of Models 1 and 2 in the downstream task

Model No.	Accuracy (5%-95% CI)		
	ABN	ACL	MEN
1 (Proposed)	0.89 (0.88-0.90)	0.90 (0.89-0.92)	0.71 (0.69-0.73)
2 (Proposed)	0.89 (0.88-0.90)	0.90 (0.89-0.92)	0.72 (0.70-0.74)

TABLE 4.5: AUC scores of Models 1 and 2 in the downstream task

Model No.	AUC (5%-95% CI)		
	ABN	ACL	MEN
1 (Proposed)	0.94 (0.93-0.95)	0.97 (0.97-0.98)	0.86 (0.84-0.88)
2 (Proposed)	0.94 (0.92-0.95)	0.97 (0.96-0.98)	0.85 (0.84-0.87)

4.5.3.2 Results of Model Pre-trained with $\mathcal{L}_{Combo}$

In Table 4.6, we present the results obtained after combining our proposed loss function with the squared Euclidean distance-based contrastive loss function. The best results from each model are written in bold, while the second best results are coloured blue.

TABLE 4.6: Accuracy and AUC scores of Models 3, 4 and 5 in the downstream task

Model No.	Accuracy			AUC		
	ABN	ACL	MEN	ABN	ACL	MEN
3 (Combo)	0.908	**0.925**	0.717	0.934	0.958	0.844
4 (Combo)	**0.925**	0.917	0.733	**0.939**	**0.964**	0.825
5 (Combo)	0.867	0.917	**0.750**	0.938	0.957	**0.848**

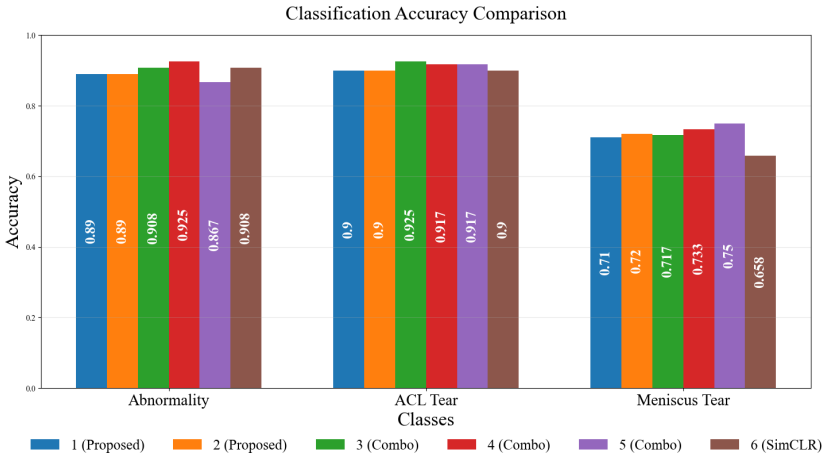
4.5.3.3 Comparison with SimCLR

The above results have been followed by comparison with results obtained from SimCLR (Chen et al., 2020a), one of the state-of-the-art self-supervised contrastive learning algorithms. The results are presented in Fig. 4.6. We can see that even though the base models 1 and 2 cannot surpass SimCLR, the combined loss $\mathcal{L}_{Combo}$ surpasses it in terms of accuracy. However, in terms of AUC score, our base models surpass the results obtained using both $\mathcal{L}_{Combo}$ and $\mathcal{L}_{SimCLR}$(Chen et al., 2020a).

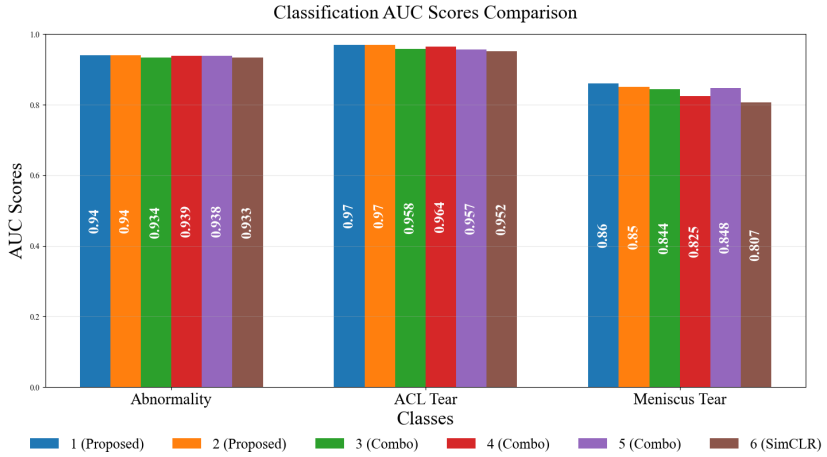
4.5.3.4 Comparison with Supervised Baseline

We also compared our proposed model and the combined model with the supervised baseline MRNet (Bien et al., 2018) model. To establish our claim that self-supervised pre-training improves performance in general, we also compare the performance of Model 6, that is the model trained with $\mathcal{L}_{SimCLR}$. All the models used in the downstream experiments were trained using only 16 slices for a single sample. The results are presented in Fig. 4.7

Self-supervised pre-training improved the performance of the model on the downstream task. The performance of the self-supervised pre-trained models surpassed the performance of the supervised model, with just 16 slices from the MR scan made available for training.



(a)



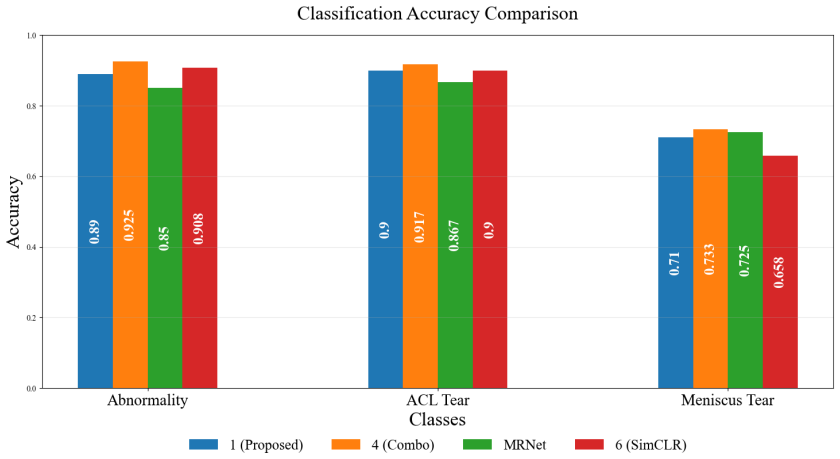
(b)

FIGURE 4.6: (a) Accuracy scores and (b) AUC scores of all models as mentioned in Table 4.2 on the downstream task.

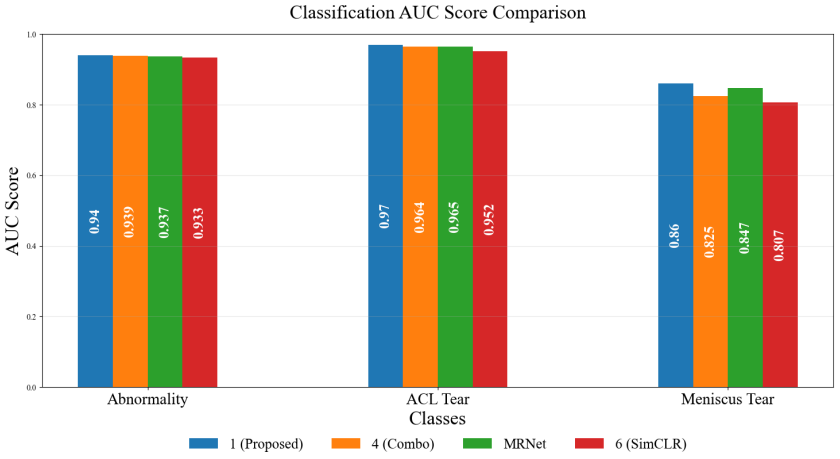
This shows that the features learnt by the self-supervised pre-training help the model greatly perform on temporally sparse data. This would help models to run both training and inference, efficiently and faster, specifically on single GPU systems.

4.5.3.5 Training Time Comparison with Supervised Baseline

As mentioned in Section 4.5.2, the total time taken for pre-training with fine-tuning on the downstream task takes approximately 7.5 hours when trained on a single NVIDIA P100 GPU. Whereas, as mentioned in the MRNet (Bien et al., 2018) paper, the training time for each out of the 9 models is 6 hours on an NVIDIA GEFORCE GTX 1070 8GB GPU. Scaling in terms of GFLOPS, the total training time of MRNet is equivalent to about



(a)



(b)

FIGURE 4.7: (a) Accuracy scores and (b) AUC scores of Models 1 ($\mathcal{L}_{Proposed}$), 4 ($\mathcal{L}_{Combo}$) and Model 6 ($\mathcal{L}_{SimCLR}$) with the supervised model MRNet. The configuration of the models is the same as in Table 4.2.

33 hours on the NVIDIA P100 GPU, which is about 4 times the required time using our method.

4.6 Conclusion

In this chapter, we proposed a novel loss function based on the binary classification of pairs in self-supervised contrastive learning. Besides that, another loss function is also used in this chapter, which is obtained by combining a squared Euclidean distance-based contrastive loss term with the proposed loss function. We have shown the mathematical

justification of the proposed loss by showing how it can be derived from the noise contrastive estimation principle. Furthermore, we have shown through experimental results that self-supervised pre-training with the proposed loss function outperforms the supervised baseline performance on the MRNet (Bien et al., 2018) dataset in terms of accuracy as well as AUC score. The lower bound analysis of the proposed loss function provides more insight into the optimization process of the models. It shows that the proposed loss works by maximizing the cosine similarity between the samples in the positive pairs and also minimizing the cosine similarity between the samples in all the pairs.

Besides providing a novel loss function, to the best of our knowledge, the framework proposed in this chapter is the first to explore the effects of self-supervised pre-training on top of ImageNet pre-training in general on visual medical data, especially MR scans. While self-supervised pre-training can help in obtaining pre-trained weights on small datasets for downstream tasks, the quality of representations improves if the quantity of data increases. Hence, to boost the pre-training quality, we show that using ImageNet pre-trained weights for self-supervised pre-training improves downstream performance. We ran all the experiments on the NVIDIA P100 GPU on Google Colab, which shows that the experiments can be reproduced easily. The success in overcoming the supervised baselines of our model can be extended to other modalities of medical and other data as well.

While in the above discussions, we assumed the imbalance factor as 1 in the proposed framework, it cannot be denied that the effect of imbalance in a batch is nullified completely. Hence, it is necessary to deal with the imbalance factor as the negative pairs play an important role in influencing the alignment and uniformity metrics in contrastive learning. To this end, in the next chapter, we take a variational perspective on the maximum likelihood estimation problem presented by the binary classification formulation of the self-supervised instance discrimination contrastive learning problem. Furthermore, we present mathematical and empirical analyses of the convergence of different contrastive and non-contrastive self-supervised frameworks.

Chapter 5

Self-Supervised Learning by Optimizing Mutual Information

5.1 Introduction

Self-supervised learning (SSL) has emerged as one of the pillars of unsupervised learning and deep learning in general. The primary aim of SSL, learning representations from unlabeled data, is fulfilled by optimizing the model's parameter values using a pre-defined task. Several innovative approaches have been proposed for the pre-training tasks which paved the way for efficient self-supervised representation learning. Both contrastive and non-contrastive learning frameworks have achieved state-of-the-art results on benchmark datasets.

In self-supervised learning, many early techniques involve solving context-based predictive pretext tasks. These tasks include geometric transformation prediction (Gidaris et al., 2018; Jing and Tian, 2018; Kumar et al., 2021), context prediction (Doersch et al., 2015; Pathak et al., 2016), jigsaw puzzle solving (Noroozi and Favaro, 2016; Kim et al., 2018; Wei et al., 2019), temporal order related tasks for videos (Misra et al., 2016; Lee et al., 2017; Fernando et al., 2017; El-Nouby et al., 2019), image colorization (Iizuka et al., 2016; Zhang et al., 2016), etc. These pretext tasks aim to learn representations invariant to transformations, context, etc. Although these tasks successfully rolled the wheels of self-supervised learning, the performances of the models pre-trained with these tasks are not at par with their supervised counterparts on the target tasks.

With the advent of CPC (van den Oord et al., 2018), the adoption of InfoNCE loss in self-supervised representation learning spread rapidly and the efficacy of instance discrimination-based frameworks on downstream tasks obtained a huge boom. Frameworks like SimCLR (Chen et al., 2020a), MoCov1 (He et al., 2020), and MoCov2 (Chen et al., 2020c) showed that it was possible to obtain performance at par with supervised learning counterparts on benchmark datasets and downstream tasks. These instance discrimination frameworks avoided the complete collapse of representations by contrasting samples in negative pairs obtained from a large batch size or memory bank. However, frameworks like BYOL (Grill et al., 2020), SimSiam (Chen and He, 2020), Barlow Twins (Zbontar et al., 2021),

and VICReg (Bardes et al., 2022a) showed that it is possible to avoid collapse without using negative pairs. It is noted that InfoNCE loss has been shown to maximize the mutual information between the samples in a positive pair (Tian et al., 2020).

In this chapter, the proposed framework intends to improve representation learning further by optimizing mutual information between the two samples in both positive and negative pairs in SSL. In other words, besides maximizing mutual information between positive pair samples, the proposed frameworks should also minimize the mutual information between negative pair samples. With this motivation, the framework proposed in this chapter is simply formulated as the task of classifying a pair as positive or negative. Different from the framework proposed in the previous chapter, this framework handles the positive and negative pairs separately to deal with the imbalance factor in each batch. While we adopt a binary classification strategy, which results in a maximum likelihood estimation problem, we discuss the base version of the proposed framework (MIOv1) from a variational problem perspective as in Pihlaja et al. (2010). Afterwards, we adopt a bottom-up approach in constructing the proposed loss function. To obtain the intermediate formulation MIOv2, we eliminated the positive-positive repulsion from the expanded mathematical expression of MIOv1. The performance is further improved by taking an upper bound of the negative-negative repulsion term in MIOv2 to increase the repulsion between the samples constituting the negative pairs. Consequently, we obtain our proposed loss MIOv3 and analytically show that the proposed loss is bounded below by the difference in the expected mutual information of the negative and the positive pairs.

In addition to that, we also attempt to analytically understand the conditions under which the instance discrimination-based contrastive SSL methods can achieve convergence. While most previous works take an information theoretic or empirical approach to understand the working principle behind their respective frameworks, in this chapter we take a novel approach by analyzing the Hessian spectrum and the Lipschitz continuity to establish the pretext under which the concept of convergence holds. This knowledge is then applied to figure out the convergence criterion using a locally satisfying Polyak-Lojasiewicz inequality. We further study the performance of the proposed framework extensively on benchmark image datasets and compare it with the state-of-the-art instance discrimination-based contrastive learning, negative-free contrastive learning, as well as, non-contrastive learning frameworks. The results reported in this chapter are also reported in Manna et al. (2021b).

The rest of the chapter is organized as follows: In Section 5.2, we discuss the notations used in the discussion and mathematical analysis in this chapter for easier understanding. Next, the motivation behind the proposed framework is described in Section 5.3. Section 5.4 describes the proposed methodology. Here, at first, the base loss function and the relation between mutual information and the proposed framework are discussed. Next, this section describes the step-by-step process to deduce the proposed loss function. Finally, the section ends with a convergence analysis of self-supervised learning frameworks. In Section 5.5, we discuss the details of the experimental configurations that are used to establish the proof of concept. This section also analyzes the performance of the proposed loss function and compares it with the other existing self-supervised algorithms, followed by ablation studies. Finally, Section 5.6 concludes the paper.

5.2 Preliminaries

Notations For the analysis, we consider the self-supervised model consisting of an encoder and a non-linear projector. Let the input, encoder, encoder output (projector input), projector, and the final feature vector (output from the projector) be denoted by x, f_θ, h, g_ψ, and z, respectively, where θ and ψ are the encoder and projector parameters, respectively. The input images $x \in \mathbb{R}^H \times \mathbb{R}^W \times \mathbb{R}^C$ when passed through the encoder f, a latent vector $h \in \mathbb{R}^F$ is obtained, where H, W, C and F are the height, width, number of channels of the images and number of channels in the encoder output h, respectively. This latent vector h gives the final feature vector $z \in \mathbb{R}^D$ when passed through the projector g, where D is the number of channels in the projector output z. The proposed loss function takes the feature vectors and outputs a scalar. To understand the flow of information we can devise the following equations

$$z = g_\psi(h) = g_\psi(f_\theta(x)) \tag{5.1}$$

5.3 Motivation

The primary objective of the self-supervised contrastive learning algorithm is to learn a mapping such that the features of the augmented versions of a sample forming a positive pair are mapped close to each other. For the samples belonging to a negative pair, the feature vectors are mapped as far as possible from each other. The primary motivation of this work is based on the fact that there are only two types of pairs in contrastive learning: positive and negative. *What if we classify the pairs as positive or negative?* This will lead the proposed contrastive learning principle to optimize the distance between any two samples in the feature space. This approach can be seen as a morphing of the InfoNCE (van den Oord et al., 2018) based contrastive learning framework into a binary classification problem. While this problem formulation is similar to the framework proposed in the previous chapter, which was derived directly from the noise contrastive estimation (Gutmann and Hyvärinen, 2012) principle, we discuss the imbalance factor present in each batch. While we can assume the imbalance factor to be 1 without any loss of generosity, the effect of imbalance on performance cannot be denied. Hence, we formulate the maximum likelihood estimation problem of binary classification from a variational perspective inspired by Pihlaja et al. (2010). In the following subsection, we will show the motivation and derivation of the formulation of the base version and further evolve it to the proposed version through intuitive and analytical discussion.

5.4 Proposed Framework

In this section, a novel loss function for contrastive learning is proposed. First, we will discuss the base loss function from which we derive our proposed loss function. Then, we will discuss the step-by-step modifications and the reason behind those to explain how we arrive at the proposed loss function in the subsequent subsections.

5.4.1 Formulation of the Vanilla Loss MIOv1

The primary motivation of the vanilla loss, MIOv1 is to classify the type of pairs in a self-supervised contrastive learning (SSCL) setting. In SSCL, we generally construct two types of pairs, positive and negative. An illustrative example of how we obtain the sets of positive and negative pairs of samples is provided in Sec. 5.2.

Let us denote the set of positive and negative pairs as $\mathcal{X}_+$ and $\mathcal{X}_-$, respectively. A pair (z_i, z_j) is assigned a binary class label k_{ij}: $k_{ij} = 1$ if $(z_i, z_j) \in \mathcal{X}_+$ and $k_{ij} = 0$ if $(z_i, z_j) \in \mathcal{X}_-$.

The objective of our SSCL framework is to calculate the posterior probabilities of the classes, given the pair of samples. In self-supervised learning, the distribution of the data, as well as the distribution of the samples are also unknown. While this objective is similar to the Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2012), the formulation of our first loss MIOv1 differs in some aspects. We will discuss the same in the following paragraphs. First, we will discuss the basic notions of NCE, and then discuss the reasons behind the deviation in MIOv1.

5.4.1.1 Binary Contrastive Learning using the notion of Noise Contrastive Estimation

In the previous chapter, we explored the binary classification formulation of self-supervised contrastive learning. Deriving from the noise contrastive estimation principle, we obtained the following loss function as given in Eqn. 5.2.

$$\mathcal{L} = -\frac{1}{T_p + T_n}\left[\sum_{(z_i,z_j)\in\mathcal{X}_+} \ln \phi((z_i, z_j); \theta) - \sum_{(z_k,z_l)\in\mathcal{X}_-} \ln(1 - \phi((z_k, z_l); \theta)\right] \tag{5.2}$$

where T_n and T_p denote the number of negative and positive pairs as defined in the previous chapter, and $\phi((z_i, z_j); \theta)$ can be written as

$$\phi((z_i, z_j); \theta) = \frac{1}{1 + \nu \exp(-\mathcal{G}_{NC}((z_i, z_j); \theta)} \tag{5.3}$$

where $\mathcal{G}_{NC}((z_i, z_j); \theta) = \frac{1}{\tau}\left(z_i \cdot z_j^T\right) = \frac{c_{ij}}{\tau}$. c_{ij} denotes the cosine similarity between z_i and z_j and is the SSL equivalent to logit values in non-parametric logistic regression. τ is the temperature hyper-parameter.

In self-supervised learning, we need to note the following two points: (1) we can always sample a batch with an equal number of positive and negative pairs, and (2) we use a non-parametric softmax / sigmoid classifier. Furthermore, the value of the ν $(= 2N - 2)$ is dependent on the batch size. Thus, for a large batch size, from Eqn. 4.11, we can say that, $P(k = 1|(z_i, z_j); \theta) \to 0$ and $P(k = 0|(z_i, z_j); \theta) \to 1$. Consequently, from Eqn. 5.2, $\mathcal{L} \to$

$-\infty$. Hence, the above interpretation of noise contrastive estimation is not entirely valid for self-supervised contrastive learning. To make the posterior probabilities independent of the batch size, we assume that the prior probabilities $P(k=1) = P(k=0) = 0.5$, that is, $\nu = 1$. However, we still need to ensure that the contribution of the positive and negative terms in the loss is equal. Otherwise, the effect of imbalance may have adverse effects on the learning process. In the next subsection, we discuss how to deal with the imbalance effect without relying on the batch size in the pre-training stage.

5.4.1.2 Why does the formulation of MIOv1 differ from Binary Cross Entropy Loss?

As previously mentioned, the primary objective of MIOv1 is to classify the type of pairs, positive or negative in self-supervised contrastive learning. This provides us with a binary logistic regression problem which requires estimating the maximum likelihood estimator (MLE). As explained in Pihlaja et al. (2010), the objective of MLE can be expressed as a variational problem, by writing the objective functional as follows,

$$\tilde{\mathcal{J}}[f] = \int p_d \log(\exp(f)) - \int p_n \frac{\exp(f)}{p_n} \tag{5.4}$$

Taking the variational derivative with respect to f, the only stationary point is given by $p_d = \exp(f)$ or $f = \log p_d$.

Replacing logarithm and identity by $g_1(\cdot)$ and $g_2(\cdot)$, respectively, Eqn. 5.4 can be expressed as,

$$\tilde{\mathcal{J}}_g[f] = \int p_d \;\; g_1\left(\frac{\exp(f)}{p_n}\right) - \int p_n \;\; g_2\left(\frac{\exp(f)}{p_n}\right) \tag{5.5}$$

The sample version of Eqn. 5.5, can be expressed as,

$$\mathcal{J}_g(\theta) = \frac{1}{N_d}\sum_{i=1}^{N_d} g_1\left(\frac{p_m(x_i;\theta)}{p_n(x_i)}\right) - \frac{1}{N_n}\sum_{i=1}^{N_n} g_2\left(\frac{p_m(y_i;\theta)}{p_n(y_i)}\right) \tag{5.6}$$

where, $(x_1, x_2, x_3, \ldots, x_{N_d})$ and $(y_1, y_2, y_3, \ldots, y_{N_n})$ are the samples from the data and auxiliary (noise) distributions, respectively.

As $N_d \to \infty$ and $N_n \to \infty$, Eqn. 5.6 reduces to,

$$\mathcal{J}_g^{\infty}(\theta) = \int p_d \; g_1\left(\frac{p_m(x_i;\theta)}{p_n(x_i)}\right) - \int p_n \; g_2\left(\frac{p_m(y_i;\theta)}{p_n(y_i)}\right) \tag{5.7}$$

Using $g_1(q) = \log(\frac{q}{1+q})$ and $g_2(q) = \log(\frac{1}{1+q})$ in Eqn. 5.6, and rearranging, we get,

$$\begin{aligned}\mathcal{J}_{NC}(\theta) &= \int p_d \log\left(\frac{1}{1+\exp\left(-\log\frac{p_n}{p_m(\theta)}\right)}\right) + \int p_n \log\left(\frac{1}{1+\exp\left(-\log\frac{p_m(\theta)}{p_n}\right)}\right) \\ &= \int p_d \log\left(\frac{1}{1+\exp\left(-\log\frac{p_n}{p_m(\theta)}\right)}\right) + \int p_n \log\left(1-\frac{1}{1+\exp\left(-\log\frac{p_n}{p_m(\theta)}\right)}\right)\end{aligned} \tag{5.8}$$

Hence, this objective function can be related to the log-likelihood in a nonlinear logistic regression model which discriminates the observed sample of p_d from the noise sample of the auxiliary density p_n, which is the very objective of MIOv1. In simple terms, we separately take the average of the likelihood terms of the positive and negative pairs, following Eqn. 5.6 and 5.7. The resulting form of MIOv1 is thus similar to $\mathcal{J}_g(\theta)$ in Eqn. 5.6. That is,

$$\mathcal{L}_{v1} = -\frac{1}{T_p}\sum_{(z_i,z_j)\in\mathcal{X}_+} \ln\phi((z_i,z_j);\theta) - \frac{1}{T_n}\sum_{(z_k,z_l)\in\mathcal{X}_-} \ln(1-\phi((z_k,z_l);\theta) \tag{5.9}$$

where T_n and T_p denote the number of negative and positive pairs as defined in the previous chapter, and $\phi((z_i,z_j);\theta)$ can be written as

$$\phi((z_i,z_j);\theta) = \frac{1}{1+\exp(-\mathcal{G}_{MIO}((z_i,z_j);\theta)} \tag{5.10}$$

where $\mathcal{G}_{MIO}((z_i,z_j);\theta)$ gives the logit value for non-parametric sigmoid classifier (logistic regression), that is, $\mathcal{G}_{MIO}((z_i,z_j);\theta) = \frac{1}{\tau}\left(z_i \cdot z_j^T\right) = \frac{c_{ij}}{\tau}$. c_{ij} denotes the cosine similarity between z_i and z_j and is the SSL equivalent to logit values in non-parametric logistic regression. τ is the temperature hyper-parameter.

A different perspective: If we use Eqn. 5.2 for MIOv1, where we have already assumed $\nu = 1$, we are causing the contribution of the positive and negative terms to the loss to be imbalanced. On the other hand, using differential averaging for the likelihood of the positive and negative terms, we cause the contribution of the respective likelihood to be equal to each other and give the virtual notion of a single positive and negative pair being used in the loss function. Therefore, the differential averaging compensates for the assumption $\nu = 1$. This approach of cost-sensitive learning is often used for learning on imbalanced data using neural networks (He and Garcia, 2009). Although, for large batch sizes, that is, $T_p, T_n \to \infty$, the scenario of imbalanced sampling no longer holds, and the imbalance factor $\nu \to 1$.

5.4.1.3 Empirical Formulation of the Vanilla Loss MIOv1

The form of MIOv1 can be further simplified for empirical analysis by substituting Equation (5.10) in Equation (5.9) and can be written as given below:

$$\mathcal{L}_{v1} = - \mathop{\mathbb{E}}_{(x_i,x_j)\sim p_+} \left[\ln \left(\frac{1}{1+e^{-\frac{c_{ij}}{\tau}}} \right) \right] - \mathop{\mathbb{E}}_{(x_k,x_l)\sim p_-} \left[\ln \left(1 - \frac{1}{1+e^{-\frac{c_{kl}}{\tau}}} \right) \right] \tag{5.11}$$

where c_{ij} is the cosine similarity between two feature vectors z_i and z_j obtained by passing x_i and x_j through the encoder and the projector. p_+ and p_- are the distribution of positive pairs and negative pairs on $\mathbb{R}^n \times \mathbb{R}^n$, respectively and τ is the temperature parameter.

Considering $\mathcal{X}_+$ and $\mathcal{X}_-$ as the sets of positive and negative pairs sampled from the distributions of positive and negative pairs, p_+ and p_-, respectively, we can rewrite $\mathcal{L}_{v1}$ as,

$$\mathcal{L}_{v1} = - \frac{1}{T_p} \sum_{(x_i,x_j)\in\mathcal{X}_+} \ln \left(\frac{1}{1+e^{-\frac{c_{ij}}{\tau}}} \right) - \frac{1}{T_n} \sum_{(x_k,x_l)\in\mathcal{X}_-} \ln \left(1 - \frac{1}{1+e^{-\frac{c_{kl}}{\tau}}} \right) \tag{5.12}$$

Following the pair formation strategy described in Section 5.2, the MIOv1 loss can be expressed as follows:

$$\begin{aligned} \mathcal{L}_{v1} &= - \frac{1}{T_p} \sum_{n=1}^{N} \left[\ln \left(\frac{1}{1+e^{-\frac{c_{nn'}}{\tau}}} \right) + \ln \left(\frac{1}{1+e^{-\frac{c_{n'n}}{\tau}}} \right) \right] \\ &\quad - \frac{1}{T_n} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m\neq n,n'}}^{2N} \ln \left(1 - \frac{1}{1+e^{-\frac{c_{nm}}{\tau}}} \right) \\ &= - \frac{1}{N} \sum_{n=1}^{N} \ln \left(\frac{1}{1+e^{-\frac{c_{nn'}}{\tau}}} \right) - \frac{1}{T_n} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m\neq n,n'}}^{2N} \ln \left(1 - \frac{1}{1+e^{-\frac{c_{nm}}{\tau}}} \right) \end{aligned} \tag{5.13}$$

where $n' = n + N$, $T_p = 2N$ and $T_n = 4N^2 - 4N$ as defined in the previous chapter. We can deduce a small relation between T_p and T_n which can be stated as $T_n = T_p^2 - 2T_p$.

5.4.2 Effect of Removing a Positive-Positive Repulsion

In this subsection, we will discuss and go through the first step on our path to obtain the proposed loss function MIOv3 from MIOv1 by formulating our intermediate loss formulation MIOv2. We take an empirical analysis approach to understand the significance and

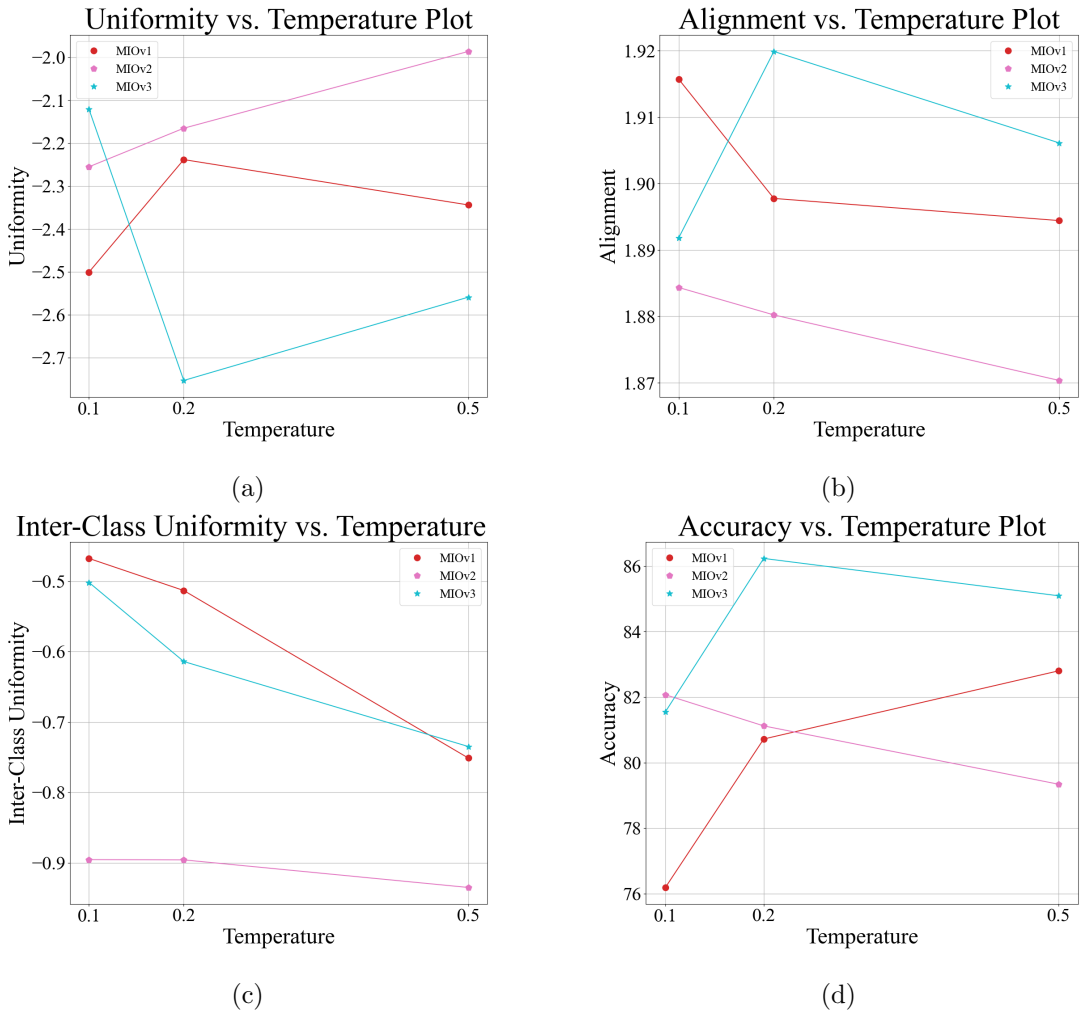


FIGURE 5.1: (a) Uniformity vs. Temperature, (b) Alignment vs. Temperature plot (c) Inter-class Uniformity vs Temperature, and (d) Accuracy vs Temperature plot at temperatures $\tau \in \{0.1, 0.2, 0.5\}$ for MIOv1, MIOv2 and MIOv3 on the CIFAR10 dataset (Krizhevsky, 2009). We did not explore temperature values above 0.5 as no improvement in performance was observed (Sec. 5.5.4.1).

working principle of MIOv2. To this end, we can expand Equation (5.13), to get,

$$\begin{aligned}\mathcal{L}_{v1} &= -\frac{1}{N}\sum_{n=1}^{N}\ln\left(\frac{1}{1+e^{-\frac{c_{nn'}}{\tau}}}\right) - \frac{1}{T_n}\sum_{n=1}^{2N}\sum_{\substack{m=1\\m\neq n,n'}}^{2N}\ln\left(1-\frac{1}{1+e^{-\frac{c_{nm}}{\tau}}}\right)\\ &= -\frac{1}{N}\sum_{n=1}^{N}\frac{c_{nn'}}{\tau} + \frac{1}{N}\sum_{n=1}^{N}\ln\left(1+e^{\frac{c_{nn'}}{\tau}}\right) + \frac{1}{T_n}\sum_{n=1}^{2N}\sum_{\substack{m=1\\m\neq n,n'}}^{2N}\ln\left(1+e^{\frac{c_{nm}}{\tau}}\right)\end{aligned} \tag{5.14}$$

where $n = n + N$ and T_n bear the same meaning as in Equation (5.13). In the Equation (5.14), we see that minimizing the loss $\mathcal{L}_{v1}$, minimizes the second term. This means that the terms $c_{nn'}$ and c_{nm} are also minimized. However, $c_{nn'}$ being the cosine similarity

of the samples in a positive pair should be maximized to +1. We see that a repulsive force will take effect on the samples in the positive pair due to the minimization of the second term in the last line of Equation 5.14. Elimination of this repulsive force should improve the performance and result in faster convergence in the optimization process. To this end, we arrive at our second loss as mentioned in Equation 5.15, which we term as MIOv2.

$$\mathcal{L}_{v2} = -\frac{1}{N}\sum_{n=1}^{N}\frac{c_{nn'}}{\tau} + \frac{1}{T_n}\sum_{n=1}^{2N}\sum_{\substack{m=1 \\ m\neq n,n'}}^{2N} \ln\left(1+e^{\frac{c_{nm}}{\tau}}\right) \tag{5.15}$$

The above equation can also be written as,

$$\mathcal{L}_{v2} = -\mathbb{E}_{(x_i,x_j)\sim p_+}\left[\frac{c_{ij}}{\tau}\right] + \mathbb{E}_{(x_k,x_l)\sim p_-}\left[\ln\left(1+e^{\frac{c_{kl}}{\tau}}\right)\right] \tag{5.16}$$

where p_+ and p_- denote the same quantities as in Equation (5.13).

To analyze the relation between the loss functions, we will first expand the expression for $\mathcal{L}_{v1}$, as follows,

$$\begin{aligned}\mathcal{L}_{v1} =& -\mathop{\mathbb{E}}_{(x_i,x_j)\sim p_+}\left[ln\left(\frac{1}{1+e^{-\frac{c_{ij}}{\tau}}}\right)\right] - \mathop{\mathbb{E}}_{(x_k,x_l)\sim p_-}\left[ln\left(1-\frac{1}{1+e^{-\frac{c_{kl}}{\tau}}}\right)\right] \\ =& -\mathbb{E}_{(x_i,x_j)\sim p_+}\left[\frac{c_{ij}}{\tau}\right] + \mathbb{E}_{(x_i,x_j)\sim p_+}\left[ln\left(1+e^{\frac{c_{ij}}{\tau}}\right)\right] + \mathbb{E}_{(x_k,x_l)\sim p_-}\left[ln\left(1+e^{\frac{c_{kl}}{\tau}}\right)\right] \\ =& \mathcal{L}_{v2} + \mathbb{E}_{(x_i,x_j)\sim p_+}\left[ln\left(1+e^{\frac{c_{ij}}{\tau}}\right)\right]\end{aligned} \tag{5.17}$$

Now, to analyze the phenomenon behind the difference in performance between MIOv1 and MIOv2 at different temperatures, we will look at how the loss functions behave. Let us consider two cases, (1) $c_{ij} > 0$, and (2) $c_{ij} \leq 0$. Without loss of generality, we can assume that $\tau > 0$. Minimizing the term $\mathbb{E}_{(x_i,x_j)\sim p_+}\left[ln\left(1+e^{\frac{c_{ij}}{\tau}}\right)\right]$ of Eqn. 5.17 increases the repulsion between the samples constituting the positive pair. We will use the notation $\mathcal{R}_{pp}$ to denote this term from here onwards. Now, for Case (1), as temperature τ increases, the magnitude of $\ln(1+e^{\frac{c_{ij}}{\tau}})$ decreases. Hence, $\mathcal{R}_{pp}$ decreases, and the repulsive force acting on the samples in the positive pairs is reduced. From Figure 5.1, we can observe that for both MIOv1 and MIOv2, as the temperature increases initially (from $\tau = 0.1$ to $\tau = 0.2$) the increase in alignment and decrease in inter-class uniformity indicates that the samples in each cluster move close to each other, and hence the rise in uniformity. This is primarily due to the effect of increasing temperature on $\mathcal{R}_{pp}$ in MIOv1. However, when τ decreases, the magnitude of $\ln(1+e^{\frac{c_{ij}}{\tau}})$ increases, consequently it increases $\mathcal{R}_{pp}$ as well as the repulsion between the samples in the positive pairs. This effect is detrimental to the performance, as the samples in the positive pairs are mapped far apart. For Case (2), the variation of $\mathcal{R}_{pp}$ with temperature will be inverted, that is, with decreasing temperature, the value of $\mathcal{R}_{pp}$ will decrease, and vice versa.

Without $\mathcal{R}_{pp}$ in MIOv2, the repulsion between the samples in the positive pair vanishes. Hence, intuitively MIOv2 should optimize better than MIOv1. At low temperatures, the magnitude of $\mathcal{R}_{pp}$ in MIOv1 increases, preventing samples in positive pairs from being mapped close to each other. However, at high temperatures, the magnitude of $\mathcal{R}_{pp}$ decreases. Consequently, the difference between MIOv1 and MIOv2 is reduced. In some cases, as empirically observed, MIOv1 outperforms MIOv2 at higher temperatures. This is primarily due to the absence of $\mathcal{R}_{pp}$.

Thus, without $\mathcal{R}_{pp}$ (MIOv2), the parameters are better optimized at lower temperatures than with $\mathcal{R}_{pp}$ (MIOv1), whereas the reverse is true at higher temperatures, as evident from Table 5.8.

5.4.3 Optimization of Negative Pair Repulsion and its Results

In MIOv2, we eliminated the positive-positive repulsion. One notable issue with lower temperatures is the instability that it can bring along as the magnitude of the gradients also increases. To maintain stability we need to tread at higher temperatures. However, we also need to maintain uniformity at higher temperatures. However, to further improve performance without positive-negative pair coupling, we start by looking at the second term of MIOv2, that is, $\mathbb{E}_{(x_k,x_l)\sim p_-}\left[\ln\left(1+e^{\frac{c_{kl}}{\tau}}\right)\right]$. We denote this term by $\mathcal{R}_{nn}$. To improve uniformity we need to increase the repulsion between samples in negative pairs further. We achieve this by incorporating the upper bound of the term mentioned above in MIOv2, resulting in our final proposed loss function, MIOv3.

To arrive at our final loss function, we follow some mathematically justifiable steps. Using Mean Value Theorem (Serret, 1868), there exists $\xi \in (0, x)$, such that,

$$ln(1+x) = ln(1+x) - ln(1) = x \cdot \left[\frac{\partial ln(1+x)}{\partial x}\right]_{x=\xi} = x \cdot \frac{1}{1+\xi} \leq x \tag{5.18}$$

Using the above relation in $ln\left(1+e^{\frac{c_{kl}}{\tau}}\right)$ from Equation (5.16), we get,

$$ln\left(1+e^{\frac{c_{kl}}{\tau}}\right) \leq e^{\frac{c_{kl}}{\tau}} \tag{5.19}$$

Replacing $ln\left(1+e^{\frac{c_{kl}}{\tau}}\right)$ by $e^{\frac{c_{kl}}{\tau}}$ in Equation (5.16), we get,

$$\begin{aligned}\mathcal{L}_{v3} &= -\mathbb{E}_{(x_i,x_j)\sim p_+}\left[\frac{c_{ij}}{\tau}\right] + \mathbb{E}_{(x_k,x_l)\sim p_-}\left[e^{\frac{c_{kl}}{\tau}}\right] \\ &= -\sum_{(x_i,x_j)\in\mathcal{X}_+}\left[\frac{c_{ij}}{\tau}\right] + \sum_{(x_k,x_l)\in\mathcal{X}_-}\left[e^{\frac{c_{kl}}{\tau}}\right]\end{aligned} \tag{5.20}$$

We can rewrite the above equation as,

$$\mathcal{L}_{v3} = -\frac{1}{N}\sum_{n=1}^{N}\frac{c_{nn'}}{\tau} + \frac{1}{T_n}\sum_{n=1}^{2N}\sum_{\substack{m=1 \\ m\neq n,n'}}^{2N} e^{\frac{c_{nm}}{\tau}} \tag{5.21}$$

where $n' = n + N$, $T_n = 2N(2N-2)$ for a batch of size N. $\mathcal{X}_+$ and $\mathcal{X}_-$ are sets of positive and negative pairs of samples obtained from the distribution of positive and negative pairs, p_+ and p_-, respectively. We call this version of the loss as MIOv3, that is, $\mathcal{L}_{v3}(g_\psi(f_\theta(X)))$, which is our final proposed loss function. Pre-training a ResNet50 (ResNet18) model on ImageNet100 (CIFAR10) using the same hyper-parameters configuration as MIOv1 and MIOv2.We observe that the performance improves considerably and even surpasses the contemporary contrastive learning frameworks on ImageNet100 (CIFAR10).

In Sec. 5.4.2, we already discussed the two cases, for which analyzed the behaviour of MIOv1 and MIOv2. With increasing temperature τ, the magnitude of the term $\mathcal{R}_{nn}$ decreases or increases depending on the cosine similarity of the samples in the concerned pair.

Let us denote the upper bound of $\mathcal{R}_{nn}$ by $\mathcal{O}_{\mathcal{R}nn}$. Now, using $\mathcal{O}_{\mathcal{R}nn}$ in place of $\mathcal{R}_{nn}$ increases the repulsion between the samples in the negative pairs. This effect helps in maintaining the uniformity of the samples, thereby preventing the collapse observed in MIOv2 at high temperatures. However, with the decrease in temperature, for the cases of $c_{ij} > 0$, $\mathcal{O}_{\mathcal{R}nn}$ grows faster than $\mathcal{R}_{nn}$. This results in an exponential increase in the repulsion between the false negative pairs. Consequently, alignment (Wang and Isola, 2020) of samples is hindered. For $c_{ij} < 0$, we observe that $\mathcal{O}_{\mathcal{R}nn} \to \mathcal{R}_{nn}$. Hence, the effect is not so evident in this case.

5.4.4 Relation of Proposed Loss MIOv3 and Mutual Information

In this subsection, we are going to derive the relationship between the MIOv3 loss function and mutual information (Shannon, 1948; Cover and Thomas, 2006; McAllester and Stratos, 2020) between the samples in a pair. The final expression of the lower bound of the MIOv3 loss function will allow us to visualize the optimization process intuitively.

Let us define the class-conditional probability densities as follows,

$$\begin{aligned} p((z_i, z_j)|k=1) &= p_m((z_i, z_j);\theta) = P_+^{i,j} \\ p((z_i, z_j)|k=0) &= p_n((z_i, z_j)) = P_-^{i,j} \end{aligned} \tag{5.22}$$

Here, $P_+^{i,j}$ is the probability of obtaining the pair (z_i, z_j) given the sample is drawn from the positive pair distribution, i.e. $k = 1$, and $P_-^{i,j}$ is the probability of obtaining the pair (z_i, z_j) given the sample is drawn from the positive pair distribution, i.e. $k = 0$. $p_m(\cdot;\theta)$ and $p_n(\cdot)$ are the normalized parameterized data distribution and noise distribution, respectively, as already discussed in Section 4.2.

In the context of the framework proposed in this chapter, we consider the imbalance factor $\nu = 1$, as already discussed in Section 5.4.1. Now, the probability of the pair (z_i, z_j) being a positive pair in a binary classification setting can be expressed as:

$$P(k=1|(z_i,z_j)) = \frac{P(k=1)P_+^{i,j}}{P(k=1)P_+^{i,j} + P(k=0)P_-^{i,j}} = \frac{P_+^{i,j}}{P_+^{i,j} + P_-^{i,j}} \tag{5.23}$$

where $P(k=1)$ and $P(k=0)$ are the class prior probabilities and $P(k=1) = P(k=0)$. The complete analysis behind the reason for considering $P(k=1) = P(k=0)$ is given in detail in Section 5.4.1.

Considering $P_Z(z_i)$ as the probability of obtaining z_i from the distribution p_Z over all possible transformed samples of z and $P_{Z,Z}(z_i, z_j)$ as the probability of obtaining (z_i, z_j) from the joint distribution $p_{Z,Z}$, we deduce the following relations. When considering (z_i, z_j) as a positive pair, the parent sample z from which we obtain a positive pair is not observed. Hence, we cannot consider z_i and z_j as independent (Koller and Friedman, 2009). In Figure 5.2, for example, the positive transformed pair (z_1, z_2) is obtained from the same sample z. Thus, $P_+^{i,j}$ is equal to the probability $P_{Z,Z}(z_i, z_j)$. Again, when considering (z_i, z_j) as a negative pair, there will be no dependency between the two samples, as any two samples can be paired to form a negative pair. For example, (z_1, z_3) or (z_2, z_4) in Figure 5.2. Thus, z_i and z_j can be considered independent and $P_-^{i,j}$ can be considered as the product of $P_Z(z_i)$ and $P_Z(z_j)$.

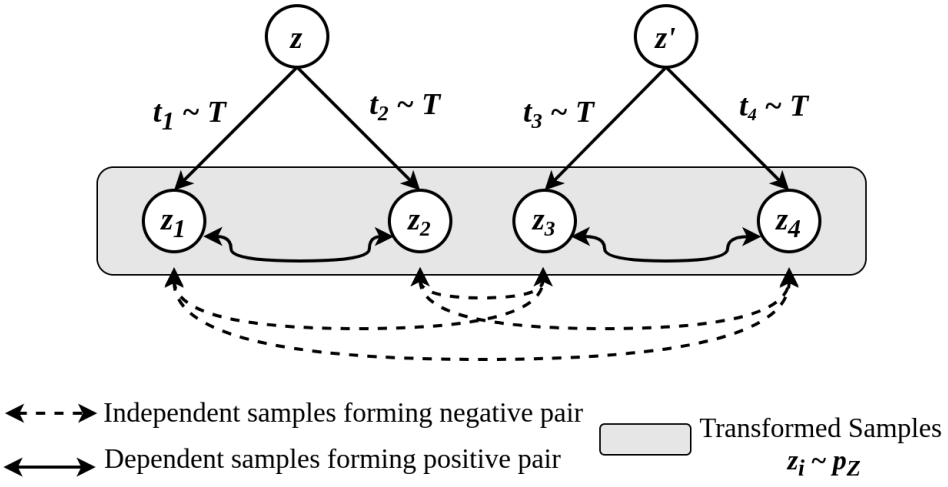


FIGURE 5.2: Graphical Model (Koller and Friedman, 2009) showing the dependence between two samples in a positive pair and the independence between two samples forming a negative pair. Here, z and z' are two different samples in a dataset. t_1, t_2, t_3, t_4 are randomly chosen transformations from the distribution T. z_1 and z_2 are obtained by applying t_1 and t_2 on z. z_3 and z_4 are obtained by applying t_3 and t_4 on z'.

Therefore, using the same idea, Equation (5.23) can be expanded as follows:

$$P(k=1|(z_i,z_j)) = \frac{p_{Z,Z}(z_i,z_j)}{p_{Z,Z}(z_i,z_j)+p_Z(z_i)p_Z(z_j)} = \frac{\frac{p_{Z,Z}(z_i,z_j)}{p_Z(z_i)p_Z(z_j)}}{1+\frac{p_{Z,Z}(z_i,z_j)}{p_Z(z_i)p_Z(z_j)}} \tag{5.24}$$

Let us define the scoring function

$$\phi_s(z_i,z_j) = e^{c_{ij}} \tag{5.25}$$

where c_{ij} is the cosine similarity between z_i and z_j.

We can also express $P(k=1|(z_i,z_j))$ in terms of the score function $\phi_s(z_i,z_j)$ as follows,

$$P(k=1|(z_i,z_j)) = \frac{1}{1+e^{-c_{ij}}} = \frac{e^{c_{ij}}}{1+e^{c_{ij}}} = \frac{\phi_s(z_i,z_j)}{1+\phi_s(z_i,z_j)} \tag{5.26}$$

Thus, comparing Equation (5.24) and (5.26), we get,

$$\phi_s(z_i,z_j) = e^{c_{ij}} = \frac{p_{Z,Z}(z_i,z_j)}{p_Z(z_i)p_Z(z_j)} \tag{5.27}$$

Putting Equation (5.27) in Equation (5.21), we get,

$$\begin{aligned}
\mathcal{L}_{v3} &= -\mathop{\mathbb{E}}_{(x_i,x_j)\sim p_+}\left[log\left(\frac{p_{Z,Z}(z_i,z_j)}{p_Z(z_i)p_Z(z_j)}\right)\right] + \mathop{\mathbb{E}}_{(x_k,x_l)\sim p_-}\left[\frac{p_{Z,Z}(z_k,z_l)}{p_Z(z_k)p_Z(z_l)}\right] \\
&= -\mathop{\mathcal{I}}_{(z_i,z_j)\sim p_+}(z_i,z_j) + \mathop{\mathbb{E}}_{(x_k,x_l)\sim p_-}\left[\frac{p_{Z,Z}(z_k,z_l)}{p_Z(z_k)p_Z(z_l)}\right] \\
&\geq -\mathop{\mathcal{I}}_{(z_i,z_j)\sim p_+}(z_i,z_j) + \mathop{\mathbb{E}}_{(x_k,x_l)\sim p_-}\left[log\left(1+\frac{p_{Z,Z}(z_k,z_l)}{p_Z(z_k)p_Z(z_l)}\right)\right] \\
&\geq -\mathop{\mathcal{I}}_{(z_i,z_j)\sim p_+}(z_i,z_j) + \mathop{\mathbb{E}}_{(x_k,x_l)\sim p_-}\left[log\left(\frac{p_{Z,Z}(z_k,z_l)}{p_Z(z_k)p_Z(z_l)}\right)\right] \\
&\geq -\mathop{\mathcal{I}}_{(z_i,z_j)\sim p_+}(z_i,z_j) + \mathop{\mathcal{I}}_{(z_k,z_l)\sim p_-}(z_k,z_l)
\end{aligned} \tag{5.28}$$

From Equation (5.28), we can infer that the proposed loss function $\mathcal{L}_{v3}$ works by maximizing the mutual information between the samples in a positive pair (z_i,z_j). It also minimizes the mutual information between the samples in a negative pair (z_k,z_l). The above derivation shows that the motivation of the proposed framework is justified when optimizing the MIOv3 loss for self-supervised representation learning.

5.5 Experimental Details, Results and Analysis

In this section, first, we will discuss the datasets used for our experiments, and then the experimental configuration used for pre-training the models. We also present the accuracies of the proposed framework on the mentioned datasets and compare them with the state-of-the-art algorithms. In our experiments, we have used natural image datasets for pre-training. This allows us to transfer the representations learnt in the pre-training task to a variety of medical datasets. Another reason for not directly using medical datasets for pre-training is the unavailability of a large medical image dataset with content as diverse as natural image datasets like ImageNet. This limits the generalization capability of the models pre-trained on the small medical image datasets, and subsequently limits the performance on the downstream task. The transferability of the pre-trained representations is also limited due to the absence of diversified features in a single medical image dataset. However, we can evaluate the quality of representations learnt from the natural image datasets in the pre-training step by using transfer learning tasks on both medical and natural image datasets with appropriate fine-tuning of the parameters. We compare the performance on transfer learning tasks with both SSL and supervised learning baselines.

5.5.1 Datasets

We use four popular datasets to conduct the experiments, namely, CIFAR-10, STL-10, CIFAR-100, Tiny ImageNet, ImageNet100 and ImageNet1K. The dimensions of images in CIFAR-10, STL-10, CIFAR-100, Tiny ImageNet, ImageNet100 and ImageNet1K are 32×32, 96×96, 32×32, 64×64, 256×256 and 256×256, respectively. The details of the distribution of the training and test sets are given in Table 5.1. To test the generalizability of the framework we use medical data along with natural image data for our experiment in this chapter.

TABLE 5.1: Training and Test images distribution in different datasets

Dataset	No. of classes	Images		Image Dimensions
		Training	Test	
CIFAR-10	10	50000	10000	32×32
CIFAR-100	100	50000	10000	32×32
STL-10	10	5000	8000	96×96
Tiny Image Net	200	100000	10000	64×64
ImageNet100	100	130000	5000	256×256
ImageNet1K	1000	1.2M	50000	256×256

5.5.2 Implementation Details

In this section, we mention the configuration of the best-performing models for the frameworks proposed in this chapter and also for the frameworks used for comparison. The frameworks were implemented using the lightly-ai (Susmelj et al., 2020) library. For the experiments on ImageNet1K and ImageNet100 datasets, we used a ResNet50 (He et al.,

2016) backbone for all our experiments. The network parameters were optimized using a LARS optimizer with the square root learning rate scaling scheme as described in the SimCLR (Chen et al., 2020a) paper. For all our experiments, we used a batch size of 256. The pre-training and the downstream tasks were run on a single 24GB NVIDIA A5000 GPU using the lightly-ai (Susmelj et al., 2020) library. To ensure faster training and prevent out-of-memory issues, we adopted automatic mixed precision (AMP) training. The time taken for pre-training on the ImageNet100 and ImageNet1K datasets is about 36 hours and 170 hours, respectively.

For the small-scale benchmarks, all the models were trained using ResNet-18 with a batch size of 128. The respective loss functions of the self-supervised models were optimized using an SGD optimizer with a learning rate of 0.06 for CIFAR10 and CIFAR100, and a learning rate of 0.03 for STL-10 and Tiny-ImageNet. The models were pre-trained for short training periods of 200 epochs only.

We decayed the learning rate following a cosine annealing schedule. The value of weight decay used is 5×10^{-4}. The ResNet architecture is modified as mentioned in Chen et al. (2020a) only for CIFAR10 and CIFAR100 datasets as the image dimensions are 32×32.

For MIOv1, MIOv2, and MIOv3, we used a temperature value of 0.2. Whereas for SimCLR (Chen et al., 2020a), DCL (Yeh et al., 2022), and DCLW (Yeh et al., 2022), we used a temperature of 0.1 as recommended in the paper Yeh et al. (2022). For MoCov2 (Chen et al., 2020c), we used a temperature value of 0.07, as recommended in its paper. The same value of temperature hyper-parameter value does not yield the best performance for all the frameworks on a particular dataset. Hence, we use a temperature value that yields the best performance for the respective frameworks.

5.5.3 Comparative Results and Analysis

In this section, we present the comparative results of the proposed framework on small-scale datasets (CIFAR-10, CIFAR-100, STL-10, Tiny-ImageNet) in Sec. 5.5.3.1, and large-scale datasets (ImageNet-100 and ImageNet-1k) in Sec. 5.5.3.2 for both contrastive and non-contrastive algorithms. We also present comparative results on transfer learning tasks on both medical and natural image datasets to evaluate the quality of learned representations.

5.5.3.1 Results on Small-Scale Datasets

In this subsection, we present the results of frameworks with MIOv1, MIOv2, and the proposed MIOv3 loss function along with the contrastive frameworks SimCLR, MoCOv2, SimCLR+DCL, SimCLR+DCLW and the non-contrastive frameworks BYOL, Barlow Twins. All the frameworks were trained and evaluated using a kNN classifier with $k = 200$ on four small-scale datasets as mentioned in Sec. 5.5.1. The Top-1 200-NN accuracy values are given in Table 5.2.

TABLE 5.2: Top-1 200-NN classification accuracy on CIFAR-10, CIFAR-100, STL-10 and Tiny ImageNet-200 datasets of SimCLR, MoCoV2, DCL, DCLW, Barlow Twins, BYOL, and MIOv3 frameworks. The configuration and implementation details are mentioned in Section 5.5.2.

Dataset	Methods						
	Contrastive				Non-Contrastive		Binary Contrastive
	SimCLR	MoCoV2	SimCLR +DCL	SimCLR +DCLW	Barlow Twins	BYOL	MIOv3
CIFAR-10	81.23	83.73	84.43	84.29	84.03	**86.84**	86.36
CIFAR-100	52.99	54.35	54.24	54.61	53.04	54.02	**58.18**
STL-10	75.65	75.64	74.46	75.49	73.24	75.87	**80.50**
Tiny ImageNet-200	24.64	29.41	29.23	30.54	27.42	21.21	**30.87**

5.5.3.2 Results on Large-Scale Datasets

In this subsection, we report the performance of the proposed framework and contrastive frameworks like SimCLR, MoCo, DCL, DCLW, and other frameworks like BYOL, DINO, ARB, VICReg, WMSE, ZeroCL. For the contrastive learning frameworks on ImageNet100 and ImageNet1K datasets, we report the Top-1 Linear evaluation accuracies.

Comparison with Contrastive Algorithms

We compare the performance of our proposed method with the contemporary state-of-the-art contrastive frameworks on the ImageNet100 and ImageNet1K datasets in Table 5.3. We pre-train our model for a duration of 200 and 100 epochs, respectively and observe that our proposed framework comfortably outperforms the state-of-the-art contrastive SSL frameworks on the Linear Evaluation task on the ImageNet100 and ImageNet1K datasets.

TABLE 5.3: Top-1 Linear evaluation accuracy on ImageNet100 and ImageNet1K datasets of SimCLR, MoCoV2, DCL, DCLW, and MIOv3 frameworks. The configuration and implementation details for each experiment are mentioned in Section 5.5.2.

Frameworks	Top -1 Linear Eval. Acc.	
	ImageNet100	ImageNet1K
SimCLR	75.54 (Huang et al., 2023a)	63.2 (Susmelj et al., 2020)
MoCoV2	76.80 (Huang et al., 2023a)	-
SimCLR+DCL	77.38 (Huang et al., 2023a)	65.1 (Susmelj et al., 2020)
SimCLR+DCLW	76.58 *(repro.)*	64.2 (Susmelj et al., 2020)
MIOv3 (Proposed)	**78.40**	**65.22**

TABLE 5.4: Comparison with state-of-the-art Non-contrastive SSL frameworks on ImageNet1K dataset (Here, B. Twins stands for Barlow Twins)

Framework	Temp. Scaled	Lin. Eval. Acc. Top-1	Top-5
BYOL (Grill et al., 2020)	N/A	62.4	82.7
B. Twins (Zbontar et al., 2021)	N/A	62.9	84.3
VicReg (Bardes et al., 2022a)	N/A	63.0	85.4
MIOv3 (Proposed)	✓	**65.22**	**86.57**

TABLE 5.5: Comparison of the proposed method with non-contrastive frameworks on the ImageNet100 dataset, pre-trained for a longer duration of 400 epochs. Here 'Linear Eval. Acc.' means Linear Evaluation Accuracy.

Frameworks	Proj. Dim #	Linear Eval. Acc. Top - 1	Top - 5
Barlow Twins (Zbontar et al., 2021)	2048	78.62	94.72
VICReg (Bardes et al., 2022a)	2048	79.22	95.06
ZeroICL (Zhang et al., 2022a)	256	78.02	95.61
ZeroFCL (Zhang et al., 2022a)	2048	79.32	94.94
ZeroCL (Zhang et al., 2022a)	2048	79.26	94.98
WMSE (Ermolov et al., 2021)	256	69.06	91.22
ARB (Zhang et al., 2022b)	2048	79.48	95.51
DINO (Caron et al., 2021)	256	74.84	92.92
BYOL (Grill et al., 2020)	4096	80.09	94.99
LogDet (Zhang et al., 2024a)	2048	80.38	95.45
MIOv3 (Proposed)	2048	**81.66**	**95.84**

Comparison with Non-Contrastive Algorithms

In this section, we compare the performance of the proposed framework with the state-of-the-art non-contrastive learning frameworks on both the ImageNet1k and ImageNet100 datasets. For the comparison on the ImageNet1k datasets, we use the results provided in the lightly-ai (Susmelj et al., 2020) benchmark results table as we use the lightly-ai library for our implementation. For the comparison on ImageNet1k in Table 5.4, we pre-trained the model for 100 epochs with a batch size of 256.

We compare the performance of our proposed method with the contemporary state-of-the-art non-contrastive frameworks on the ImageNet100 dataset in Table 5.5. We pre-train our model for a longer duration of 400 epochs following ZeroCL (Zhang et al., 2022a) and ARB (Zhang et al., 2022b). We observe that our proposed framework comfortably outperforms the state-of-the-art non-contrastive frameworks on the Linear Evaluation task on the ImageNet100 dataset.

5.5.3.3 Comparison of Performance in Transfer Learning Setting

It is imperative to show the quality of representations learnt by the self-supervised models on other datasets. For this purpose, we chose four medical image datasets, MURA (Rajpurkar et al., 2017), Chaoyang (Zhu et al., 2022), ISIC2016 Lesion Classification (Gutman et al., 2016), and MHIST (Wei et al., 2021) datasets. We also use three natural image datasets, Flowers (Nilsback and Zisserman, 2006, 2008), CIFAR10 and CIFAR100 (Krizhevsky, 2009).

Transfer Learning Performance on Medical Image Datasets
We encounter both multi-class and binary classification tasks in this section. The MURA, ISIC2016 and MHIST datasets consist of binary labels and can be used in binary classification tasks. The Chaoyang is a multi-class dataset and can be used for a multi-class classification task. The MURA (Rajpurkar et al., 2017) dataset is a Musculoskeletal Radiograph dataset, that is, it consists of bone X-ray images. The task of an algorithm is to classify whether an X-ray is abnormal or normal. The Chaoyang (Zhu et al., 2022) dataset, on the other hand, is histopathological, consisting of 6160 patches of Colon cancer divided into four classes - normal, serrated, adenocarcinoma, and adenoma. The ISIC2016 (Gutman et al., 2016) dataset contains skin lesion images, for 2 classes - malignant, and benign. There are only 900 and 379 images in the train and test set, respectively. The MHIST (Wei et al., 2021) dataset contains about 2K and 1K Hematoxylin and Eosin (H&E)-stained Formalin-Fixed Paraffin-Embedded (FFPE) fixed-size images of colorectal polyps in the train and test set, respectively and contains two classes Hyperplastic Polyp (HP), and Sessile Serrated Adenoma (SSA), annotated by 7 pathologists.

We used the models pre-trained on the ImageNet1K (Deng et al., 2009) dataset for 100 epochs and fine-tuned them for 50 epochs on these datasets, using an SGD optimizer. We also used class weights to mitigate the effect of imbalance in all the medical datasets, and a batch size of 128 in all the experiments. For the multiclass and binary classification tasks, learning rates of 0.1 and 1.0 were used, respectively, and a multi-step decay scheduler with a decay by a factor of 0.1 at the 30th and 40th epochs. For the MURA, ISIC2016, and MHIST datasets, we used positive class weights of 0.7097, 4.24, and 2.45, respectively. For the Chaoyang dataset, the class weights used were $1.264, 1.667, 1.0$, and 2.114 for the 4 classes.

From the results presented in Table 5.6, we can see that the proposed method outperforms SimCLR (Chen et al., 2020a) on all 4 medical image datasets. The proposed method also outperforms the contemporary state-of-the-art self-supervised contrastive learning algorithm (DCL (Yeh et al., 2022)) on 2 out of 4 datasets. It can also be seen that the performance of our proposed framework is close to the supervised baseline.

Transfer Learning Performance on Natural Image Datasets
The Flowers, CIFAR10, and CIFAR100 datasets are multi-class datasets and are used in multi-class classification tasks. The Flowers (Nilsback and Zisserman, 2006) dataset images for 17 classes, with only 80 images in each class. The CIFAR10 (Krizhevsky, 2009) and CIFAR100 (Krizhevsky, 2009) datasets have already been introduced in Sec. 5.5.1.

TABLE 5.6: Performance comparison of the proposed method (MIOv3) with contemporary self-supervised contrastive state-of-the-art methods on transfer learning tasks on medical image datasets. The results of the supervised learning baseline are also provided here for reference.

Datasets	SimCLR	DCL	MIOv3	Supervised
MURA	81.81	81.70	**82.49**	82.10 (Nauta et al., 2023b,a)
Chaoyang	83.22	83.12	**84.34**	83.50 (Galdran et al., 2023)
ISIC2016	84.70	**85.48**	85.22	85.50 (ISDIS, 2016)
MHIST	83.62	**85.26**	84.03	86.90 (Springenberg et al., 2023)

We used the same setting for fine-tuning as in the above subsection. We used class weights to mitigate the effect of imbalance in all datasets, except CIFAR10 and CIFAR100, as both are balanced datasets.

From the results presented in Table 5.7, we can see that the proposed method outperforms SimCLR (Chen et al., 2020a) on all 3 natural image datasets. The proposed method also outperforms the contemporary state-of-the-art self-supervised contrastive learning algorithm (DCL (Yeh et al., 2022)) on 3 out of 3 datasets. It can also be seen that the performance of our proposed framework is close to the supervised baseline on the natural image datasets too.

TABLE 5.7: Performance comparison of the proposed method (MIOv3) with contemporary self-supervised contrastive state-of-the-art methods on transfer learning tasks on natural image datasets. The results of the supervised learning baseline are also provided here for reference.

Datasets	SimCLR	DCL	MIOv3	Supervised
CIFAR10	96.93	97.09	**97.11**	97.50 (Grill et al., 2020)
CIFAR100	82.99	83.03	**83.77**	86.40 (Grill et al., 2020)
Flowers	93.82	94.11	**94.70**	97.60 (Grill et al., 2020)

5.5.3.4 Analysis of Convergence of Contrastive SSL Frameworks

The loss landscape of the different models in the frameworks depends on the loss function used. The function $\mathcal{L}_{v3} \circ g_\psi \circ f_\theta$ is a non-convex function of the parameter space $\mathbb{P}$. The input pair space χ is mapped to the latent space $\mathbb{R}^D$ by a function $g_\psi \circ f_\theta$ or $(g \circ f)_\mathcal{P}$, where $\mathcal{P} = \{\theta, \psi\}$ denotes a point in the parameter space $\mathbb{P}$. The paired embedding obtained

from the function $g_\psi \circ f_\theta$ or $(g \circ f)_\mathcal{P}$, in the self-supervised pre-training phase, constitutes a point in the embedding space $\mathcal{E} : \mathbb{R}^D \times \mathbb{R}^D$ and is mapped to the loss landscape $\mathbb{L}$, i.e., $\mathcal{L}_{v3} \circ (g \circ f)_\mathcal{P} : \chi \to \mathbb{L}$.

To analytically check if SSL pre-training truly converges, we need to proceed in three short steps. First, we need to calculate the Hessian of $\mathcal{L}_{v3}$ with respect to the parameters. Without loss of generality, we show the Hessian of $\mathcal{L}_{v3}$ with respect to ψ. Next, we need to check if $\mathcal{L}_{v3}$ has a L-Lipschitz continuous gradient with respect to the parameters. We show that the norm of the Hessian matrix $\mathcal{H}$ is bounded by L indirectly by showing that the composite function approximated by $\mathcal{L}_{v3} \circ g_\psi \circ f_\theta : \chi \to \mathbb{R}$ has a Lipschitz continuous gradient, under the constraint that $\sum_d h_{\theta n}^{(d)} < \infty$, and $\sum_{w \in \mathcal{P}} w < \infty$. Thus, we prove that $\mathcal{L}_{v3} \circ g_\psi \circ f_\theta$ belongs to a class of twice-differentiable continuous real-valued functions. Finally, we use the Polyak-Lojasiewicz (PL) inequality, defined in the local neighbourhood of the initialization point, to show that the SSL methods converge to local minima only under long pre-training.

Gradient of $\mathcal{L}_{v3}$ Let us rewrite the expression for $\mathcal{L}_{v3}$ again,

$$\mathcal{L}_{v3} = -\frac{1}{N}\sum_{n=1}^{N}\frac{c_{nn'}}{\tau} + \frac{1}{T_n}\sum_{n=1}^{2N}\sum_{\substack{m=1 \\ m\neq n,n'}}^{2N} e^{\frac{c_{nm}}{\tau}} \tag{5.29}$$

$$\mathcal{L}_{v1} = -\frac{1}{N}\sum_{n=1}^{N}\frac{c_{nn'}}{\tau} + \frac{1}{N}\sum_{n=1}^{N} ln\left(1 + e^{\frac{c_{nn'}}{\tau}}\right) + \frac{1}{T_n}\sum_{n=1}^{2N}\sum_{\substack{m=1 \\ m\neq n,n'}}^{2N} ln\left(1 + e^{\frac{c_{nm}}{\tau}}\right) \tag{5.30}$$

where $n' = n+N$, N is the batch size, τ is the temperature hyperparameter and c_{ij} denotes the cosine similarity between the feature vectors z_i and z_j. The feature vectors z_i and z_j are obtained by passing the input $x_i, x_j \in \chi$ through the encoder f_θ with parameters θ and the projector g_ψ with parameters ψ. Thus,

$$z_n = g_\psi\left(f_\theta\left(x_n\right)\right) \tag{5.31}$$

Without the loss of generality, we can assume that the projector consists of a single perceptron layer or we can consider a two-layered perceptron model with weights $\psi^{(1)}_{Di\times H}$ and $\psi^{(2)}_{H\times Do}$ for the two layers as a single layer with weights $\psi = \psi^{(1)}_{Di\times H} \cdot \psi^{(2)}_{H\times Do}$.

$$\begin{aligned}\psi &= \psi^{(1)}_{Di\times H} \cdot \psi^{(2)}_{H\times Do} \\ &= \begin{pmatrix} \psi^{(1)}_{1,1}, \psi^{(1)}_{1,2}, \ldots, \psi^{(1)}_{1,H} \\ \psi^{(1)}_{2,1}, \psi^{(1)}_{2,2}, \ldots, \psi^{(1)}_{2,H} \\ \vdots \\ \psi^{(1)}_{D_i,1}, \psi^{(1)}_{D_i,1}, \ldots, \psi^{(1)}_{D_i,H} \end{pmatrix} \begin{pmatrix} \psi^{(2)}_{1,1}, \psi^{(2)}_{1,2}, \ldots, \psi^{(2)}_{1,D_o} \\ \psi^{(2)}_{2,1}, \psi^{(2)}_{2,2}, \ldots, \psi^{(2)}_{2,D_o} \\ \vdots \\ \psi^{(2)}_{H,1}, \psi^{(2)}_{H,1}, \ldots, \psi^{(2)}_{H,D_o} \end{pmatrix} = \begin{pmatrix} \psi_{1,1}, \psi_{1,2}, \ldots, \psi_{1,D_o} \\ \psi_{2,1}, \psi_{2,2}, \ldots, \psi_{2,D_o} \\ \vdots \\ \psi_{D_i,1}, \psi_{D_i,1}, \ldots, \psi_{D_i,D_o} \end{pmatrix}_{D_i\times D_o}\end{aligned} \tag{5.32}$$

To get the final feature vector z_n, we multiply the output of the encoder $h_{\theta n}$ with the transpose of the weight matrix ψ. The shape of the feature vector z_n is $D_o \times 1$. Continuing, we get,

$$\begin{aligned} z_n =& (\psi_{D_i \times D_o})^T \cdot f_\theta\left(x_n\right)_{D_i \times 1} = (\psi_{D_i \times D_o})^T \cdot (h_{\theta n})_{D_i \times 1} = \psi^T_{D_o \times D_i} \cdot (h_{\theta n})_{D_i \times 1} \\ =& \begin{pmatrix} (\psi_{\forall,1})^T \\ (\psi_{\forall,2})^T \\ \cdot \\ (\psi_{\forall,k})^T \\ \cdot \\ (\psi_{\forall,D_o})^T \end{pmatrix} \cdot \begin{pmatrix} (h_\theta)_{1,1} \\ (h_\theta)_{2,1} \\ \cdot \\ (h_\theta)_{k,1} \\ \cdot \\ (h_\theta)_{D_i,1} \end{pmatrix} \end{aligned} \tag{5.33}$$

where $(\psi_{\forall,k})^T$ denotes the transposed version of the k-th column of the weight matrix ψ, or the k-th row of the matrix ψ^T. The k-th element of the feature vector z_n is obtained by

$$z_n^{(k)} = (\psi_{\forall,k})^T \cdot h_{\theta n} = h^T_{\theta n} \cdot \psi_{\forall,k} \tag{5.34}$$

Taking the derivative of $\mathcal{L}_{v3}$ with respect to a column of ψ, or a row of ψ^T, we get

$$\begin{aligned} \frac{\partial \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^T} =& -\frac{1}{N}\sum_{n=1}^{N} \frac{\partial \frac{c_{nn'}}{\tau}}{\partial (\psi_{\forall,k})^T} + \frac{1}{T_n}\sum_{n=1}^{2N}\sum_{\substack{m=1 \\ m\neq n,n'}}^{2N} \frac{\partial e^{\frac{c_{nm}}{\tau}}}{\partial (\psi_{\forall,k})^T} \\ =& -\frac{1}{N\tau}\sum_{n=1}^{N} \frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_{n'}^{(i)}}{\partial (\psi_{\forall,k})^T} + \frac{1}{T_n}\sum_{n=1}^{2N}\sum_{\substack{m=1 \\ m\neq n,n'}}^{2N} e^{\frac{c_{nm}}{\tau}} \frac{\partial \left(\frac{c_{nm}}{\tau}\right)}{\partial (\psi_{\forall,k})^T} \\ =& -\frac{1}{N\tau}\sum_{n=1}^{N} \frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_{n'}^{(i)}}{\partial (\psi_{\forall,k})^T} + \frac{1}{T_n\tau}\sum_{n=1}^{2N}\sum_{\substack{m=1 \\ m\neq n,n'}}^{2N} e^{\frac{c_{nm}}{\tau}} \frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_m^{(i)}}{\partial (\psi_{\forall,k})^T} \end{aligned} \tag{5.35}$$

We have a common expression $\frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_m^{(i)}}{\partial (\psi_{\forall,k})^T}$ (a column vector) in all the three terms in the expression for $\frac{\partial \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^T}$. So, we will evaluate it first and then continue with our derivation,

$$
\begin{aligned}
\frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_m^{(i)}}{\partial (\psi_{\forall,k})^T} &= \sum_{i=1}^{Do} \left(z_n^{(i)} \cdot \frac{\partial z_m^{(i)}}{\partial (\psi_{\forall,k})^T} + \frac{\partial z_n^{(i)}}{\partial (\psi_{\forall,k})^T} \cdot z_m^{(i)} \right) \\
&= \sum_{i=1}^{Do} \left(z_n^{(i)} \cdot \frac{\partial \left((\psi_{\forall,i})^T \cdot h_{\theta m} \right)}{\partial (\psi_{\forall,k})^T} + \frac{\partial \left((\psi_{\forall,i})^T \cdot h_{\theta n} \right)}{\partial (\psi_{\forall,k})^T} \cdot z_m^{(i)} \right) \\
&= \sum_{i=1}^{Do} \left(z_n^{(i)} \cdot \frac{\partial (\psi_{\forall,i})^T}{\partial (\psi_{\forall,k})^T} \cdot h_{\theta m} + \frac{\partial (\psi_{\forall,i})^T}{\partial (\psi_{\forall,k})^T} \cdot h_{\theta n} \cdot z_m^{(i)} \right) \\
&= \left(z_n^{(k)} \cdot \frac{\partial (\psi_{\forall,k})^T}{\partial (\psi_{\forall,k})^T} \cdot h_{\theta m} + \frac{\partial (\psi_{\forall,k})^T}{\partial (\psi_{\forall,k})^T} \cdot h_{\theta n} \cdot z_m^{(k)} \right) \\
&= z_n^{(k)} \cdot h_{\theta m} + h_{\theta n} \cdot z_m^{(k)}
\end{aligned}
\tag{5.36}
$$

We denote the expression $\sum_{i=1}^{Do} z_n^{(i)} \cdot z_m^{(i)}$ by c_{nm} and its derivative with respect to $(\psi_{\forall,k})^T$, i.e. $\frac{\partial c_{nm}}{\partial (\psi_{\forall,k})^T} = z_n^{(k)} \cdot h_{\theta m} + h_{\theta n} \cdot z_m^{(k)}$ by $A_{n,m}^{(k)}$. Dimension of $A_{n,m}^{(k)}$ and subsequently of $\frac{\partial \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^T}$ is $D_i \times 1$. That is $A_{n,m}^{(k)}$ and subsequently $\frac{\partial \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^T}$ is a ***column vector***.

Putting Eqn. 5.36 in Eqn. 5.35, we get,

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^T} &= -\frac{1}{N\tau} \sum_{n=1}^{N} \frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_{n'}^{(i)}}{\partial (\psi_{\forall,k})^T} + \frac{1}{T_n \tau} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} e^{\frac{c_{nm}}{\tau}} \frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_m^{(i)}}{\partial (\psi_{\forall,k})^T} \\
&= -\frac{1}{N\tau} \sum_{n=1}^{N} A_{n,n'}^{(k)} + \frac{1}{T_n \tau} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} e^{\frac{c_{nm}}{\tau}} A_{n,m}^{(k)}
\end{aligned}
\tag{5.37}
$$

where $n' = n + N$.

Therefore, using $n' = n + N$,

$$
\frac{\partial \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^T} = -\frac{1}{N\tau} \sum_{n=1}^{N} A_{n,n'}^{(k)} + \frac{1}{T_n \tau} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} e^{\frac{c_{nm}}{\tau}} A_{n,m}^{(k)} \tag{5.38}
$$

Hessian of $\mathcal{L}_{v3}$ We have already calculated the first derivative of $\mathcal{L}_{v3}$ with respect to the parameters ψ or $(\psi_{\forall,k})^T$. We proceed to calculate the Hessian of the loss function $\mathcal{L}_{v3}$ with respect to ψ in a similar manner to the first derivative.

Taking derivative of $\frac{\partial \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^T}$ with respect to $\psi_{\forall,l}$, l-th column of ψ, we get,

$$
\begin{aligned}
\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,l} \partial (\psi_{\forall,k})^T} &= \left[\frac{\partial}{\partial \psi_{1,l}} \frac{\partial \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^T}, \ldots, \frac{\partial}{\partial \psi_{D_i,l}} \frac{\partial \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^T} \right]^T \\
&= -\frac{1}{N\tau} \sum_{n=1}^{N} \frac{\partial A_{n,n'}^{(k)}}{\partial \psi_{\forall,l}} + \frac{1}{T_n \tau} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} \frac{\partial}{\partial \psi_{\forall,l}} \left(e^{\frac{c_{nm}}{\tau}} A_{n,m}^{(k)} \right)
\end{aligned}
\tag{5.39}
$$

where $n' = n + N$. Now, let us first calculate $\frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_m^{(i)}}{\partial \psi_{\forall,l}}$ (a row vector).

$$\begin{aligned}
\frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_m^{(i)}}{\partial \psi_{\forall,l}} =& \sum_{i=1}^{Do} \left(z_n^{(i)} \cdot \frac{\partial z_m^{(i)}}{\partial \psi_{\forall,l}} + z_m^{(i)} \cdot \frac{\partial z_n^{(i)}}{\partial \psi_{\forall,l}} \right) \\
=& \sum_{i=1}^{Do} \left(z_n^{(i)} \cdot \frac{\partial \left((\psi_{\forall,i})^T \cdot h_{\theta m} \right)}{\partial \psi_{\forall,l}} + z_m^{(i)} \cdot \frac{\partial \left((\psi_{\forall,i})^T \cdot h_{\theta n} \right)}{\partial \psi_{\forall,l}} \right) \\
=& \sum_{i=1}^{Do} \left(z_n^{(i)} \cdot \frac{\partial \left(h_{\theta m}^T \cdot \psi_{\forall,i} \right)}{\partial \psi_{\forall,l}} + z_m^{(i)} \cdot \frac{\partial \left(h_{\theta n}^T \cdot \psi_{\forall,i} \right)}{\partial \psi_{\forall,l}} \right) \\
=& \sum_{i=1}^{Do} \left(z_n^{(i)} \cdot h_{\theta m}^T \cdot \frac{\partial \psi_{\forall,i}}{\partial \psi_{\forall,l}} + z_m^{(i)} \cdot h_{\theta n}^T \cdot \frac{\partial \psi_{\forall,i}}{\partial \psi_{\forall,l}} \right) \\
=& z_n^{(l)} \cdot h_{\theta m}^T + z_m^{(l)} \cdot h_{\theta n}^T
\end{aligned} \tag{5.40}$$

Let us denote $\frac{\partial \sum_{i=1}^{Do} z_n^{(i)} \cdot z_m^{(i)}}{\partial \psi_{\forall,l}} = z_n^{(l)} \cdot h_{\theta m}^T + z_m^{(l)} \cdot h_{\theta n}^T$ by $A_{n,m}^{(l)T}$, whose dimension is $1 \times D_i$.

Now, let us separately evaluate, $\frac{\partial}{\partial \psi_{\forall,l}} \left(e^{\frac{c_{nm}}{\tau}} A_{n,m}^{(k)} \right)$ first to make our life easier.

$$\begin{aligned}
\frac{\partial}{\partial \psi_{\forall,l}} \left(e^{\frac{c_{nm}}{\tau}} A_{n,m}^{(k)} \right) &= e^{\frac{c_{nm}}{\tau}} \frac{\partial A_{n,m}^{(k)}}{\partial \psi_{\forall,l}} + A_{n,m}^{(k)} \frac{\partial}{\partial \psi_{\forall,l}} e^{\frac{c_{nm}}{\tau}} \\
&= e^{\frac{c_{nm}}{\tau}} \frac{\partial A_{n,m}^{(k)}}{\partial \psi_{\forall,l}} + \frac{1}{\tau} A_{n,m}^{(k)} e^{\frac{c_{nm}}{\tau}} \frac{\partial c_{nm}}{\partial \psi_{\forall,l}} \\
&= e^{\frac{c_{nm}}{\tau}} \frac{\partial A_{n,m}^{(k)}}{\partial \psi_{\forall,l}} + \frac{1}{\tau} e^{\frac{c_{nm}}{\tau}} \cdot A_{n,m}^{(k)} \cdot A_{n,m}^{(l)T}
\end{aligned} \tag{5.41}$$

The only thing left to calculate is $\frac{\partial A_{n,m}^{(k)}}{\partial \psi_{\forall,l}}$. The final dimension of this quantity will be $D_i \times D_i$. Let us denote this quantity by $B_{n,m}^{(l)(k)}$.

$$\begin{aligned}
\frac{\partial A_{n,m}^{(k)}}{\partial \psi_{\forall,l}} &= \frac{\partial \left(z_n^{(k)} \cdot h_{\theta m} + z_m^{(k)} \cdot h_{\theta n} \right)}{\partial \psi_{\forall,l}} = h_{\theta m} \cdot \frac{\partial z_n^{(k)}}{\partial \psi_{\forall,l}} + h_{\theta n} \cdot \frac{\partial z_m^{(k)}}{\partial \psi_{\forall,l}} \\
&= h_{\theta m} \cdot \frac{\partial \left((\psi_{\forall,k})^T \cdot h_{\theta n} \right)}{\partial \psi_{\forall,l}} + h_{\theta n} \cdot \frac{\partial \left((\psi_{\forall,k})^T \cdot h_{\theta m} \right)}{\partial \psi_{\forall,l}} \\
&= h_{\theta m} \cdot \frac{\partial \left(h_{\theta n}^T \cdot (\psi_{\forall,k}) \right)}{\partial \psi_{\forall,l}} + h_{\theta n} \cdot \frac{\partial h_{\theta m}^T \cdot \left((\psi_{\forall,k})^T \right)}{\partial \psi_{\forall,l}} \\
&= h_{\theta m} \cdot h_{\theta n}^T + h_{\theta n} \cdot h_{\theta m}^T \big|_{k=l} \text{ or } 0 \big|_{k \neq l}
\end{aligned} \tag{5.42}$$

Therefore, $B_{n,m}^{(l)(k)} = 0$ and $B_{n,m}^{(l)(l)} = h_{\theta m} \cdot h_{\theta n}^T + h_{\theta n} \cdot h_{\theta m}^T$.

Putting Eqn. 5.41 and 5.42 in Eqn. 5.39, we get,

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,l} \partial (\psi_{\forall,k})^T} =& -\frac{1}{N\tau} \sum_{n=1}^{N} \frac{\partial A_{n,n'}^{(k)}}{\partial \psi_{\forall,l}} + \frac{1}{T_n \tau} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} \frac{\partial}{\partial \psi_{\forall,l}} \left(e^{\frac{c_{nm}}{\tau}} A_{n,m}^{(k)} \right) \\ =& \frac{1}{T_n \tau^2} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} e^{\frac{c_{nm}}{\tau}} \cdot A_{n,m}^{(k)} \cdot A_{n,m}^{(l)T}\end{aligned} \tag{5.43}$$

and,

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,k} \partial (\psi_{\forall,k})^T} =& -\frac{1}{N\tau} \sum_{n=1}^{N} \frac{\partial A_{n,n'}^{(k)}}{\partial \psi_{\forall k}} + \frac{1}{T_n \tau} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} \frac{\partial}{\partial \psi_{\forall k}} \left(e^{\frac{c_{nm}}{\tau}} A_{n,m}^{(k)} \right) \\ =& -\frac{1}{N\tau} \sum_{n=1}^{N} B_{n,n'}^{(k)(k)} \\ &+ \frac{1}{T_n \tau} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} \left(e^{\frac{c_{nm}}{\tau}} B_{n,m}^{(k)(k)} + \frac{1}{\tau} e^{\frac{c_{nm}}{\tau}} \cdot A_{n,m}^{(k)} \cdot A_{n,m}^{(k)T} \right)\end{aligned} \tag{5.44}$$

We can write the two equations Eqn. 5.43 and Eqn. 5.44, in a single equation, in a general form, as follows,

$$\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,l} \partial (\psi_{\forall,k})^T} = -\frac{1}{N\tau} \sum_{n=1}^{N} B_{n,n'}^{(l)(k)} + \frac{1}{T_n \tau} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} \left(e^{\frac{c_{nm}}{\tau}} B_{n,m}^{(l)(k)} + \frac{1}{\tau} e^{\frac{c_{nm}}{\tau}} \cdot A_{n,m}^{(k)} \cdot A_{n,m}^{(l)T} \right) \tag{5.45}$$

where

$$B_{n,m}^{(l)(k)} = \begin{cases} 0, \text{ if } l = k \\ h_{\theta m} \cdot h_{\theta n}^T + h_{\theta n} \cdot h_{\theta m}^T, \text{ if } l \neq k \end{cases}$$

We took each row in the weight matrix as a single variable for ease of calculation. This results in the second derivative being a matrix. The terms $\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,k} \partial (\psi_{\forall,k})^T}$ and $\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,l} \partial (\psi_{\forall,k})^T}$ are matrices themselves. Each element in the resultant matrix corresponds to each second derivative element in $\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,k} \partial (\psi_{\forall,k})^T}$ or $\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,l} \partial (\psi_{\forall,k})^T}$, each with dimensions $D_i \times D_i$.

Expansion of $\frac{\partial^2 \mathcal{L}_{v3}}{\partial (\psi_{\forall,k})^{T^2}}$ and $\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,l} \partial (\psi_{\forall,k})^T}$

$$\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,l} \partial (\psi_{\forall,k})^T} = \begin{pmatrix} \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{1l} \partial \psi_{1k}} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{1l} \partial \psi_{Dik}} \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{il} \partial \psi_{1k}} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{il} \partial \psi_{Dik}} \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{Dil} \partial \psi_{1k}} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{Dil} \partial \psi_{Dik}} \end{pmatrix} \tag{5.46}$$

$$\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall k} \partial (\psi_{\forall,k})^T} = \begin{pmatrix} \frac{\partial^2 \mathcal{L}_{v3}}{\partial {\psi_{1k}}^2} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{1k} \partial \psi_{Dik}} \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{jk} \partial \psi_{1k}} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{jk} \partial \psi_{Dik}} \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{Dik} \partial \psi_{1k}} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial {\psi_{Dik}}^2} \end{pmatrix} \tag{5.47}$$

Essentially, the Hessian matrix $\mathcal{H}$ should be a $\mathcal{N}_\mathcal{P} \times \mathcal{N}_\mathcal{P}$ matrix, where $\mathcal{N}_\mathcal{P}$ is the number of parameters in the model whose parameters are being optimized. The expression of Hessian matrix $\mathcal{H}$ is as follows

$$\mathcal{H}(\psi) = \begin{pmatrix} \frac{\partial^2 \mathcal{L}_{v3}}{\partial {\psi_{11}}^2} & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{11} \partial \psi_{12}} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{11} \partial \psi_{D_i D_o}} \\ \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{12} \partial \psi_{11}} & \frac{\partial^2 \mathcal{L}_{v3}}{\partial {\psi_{22}}^2} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{12} \partial \psi_{D_i D_o}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{jk} \partial \psi_{11}} & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{jk} \partial \psi_{12}} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{jk} \partial \psi_{D_i D_o}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{D_i D_o} \partial \psi_{11}} & \frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{D_i D_o} \partial \psi_{12}} & \cdots & \frac{\partial^2 \mathcal{L}_{v3}}{\partial {\psi_{D_i D_o}}^2} \end{pmatrix} \tag{5.48}$$

L-Lipschitz continuous gradient of the $\mathcal{L}_{v3} \circ g_\psi \circ f_\theta$ To prove that the loss function $\mathcal{L}_{v3} \circ g_\psi \circ f_\theta$ has L-Lipschitz continuous gradient, we need to show that the spectral norm of the Hessian matrix $\mathcal{H}$ is upper bounded by L. We can also prove that,

$$\begin{aligned} &\|\nabla \mathcal{L}_{v3} \circ g_{\psi_{t+1}} \circ f_{\theta_{t+1}}(x) - \nabla \mathcal{L}_{v3} \circ g_{\psi_t} \circ f_{\theta_t}(x)\| \\ &= \|\nabla \mathcal{L}_{v3} \circ (g \circ f)_{\mathcal{P}_{t+1}}(x) - \nabla \mathcal{L}_{v3} \circ (g \circ f)_{\mathcal{P}_t}(x)\| \leq L \|\mathcal{P}_{t+1} - \mathcal{P}_t\| \end{aligned} \tag{5.49}$$

We take a single element of the Hessian matrix $\mathcal{H}$, $\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,l} \partial (\psi_{\forall,k})^T}$ and analyse it analytically. Thus,

$$\frac{\partial^2 \mathcal{L}_{v3}}{\partial \psi_{\forall,l} \partial (\psi_{\forall,k})^T} = -\frac{1}{N\tau} \sum_{n=1}^{N} B_{n,n'}^{(l)(k)} + \frac{1}{T_n \tau} \sum_{n=1}^{2N} \sum_{\substack{m=1 \\ m \neq n,n'}}^{2N} \left(e^{\frac{c_{nm}}{\tau}} B_{n,m}^{(l)(k)} + \frac{1}{\tau} e^{\frac{c_{nm}}{\tau}} \cdot A_{n,m}^{(k)} \cdot A_{n,m}^{(l)T} \right) \tag{5.50}$$

Since the exponential terms exist in the Hessian terms, we can say that the function $(\mathcal{L}_{v3} \circ g_\psi \circ f_\theta)(x)$ belongs to the class of C^∞ functions, provided $\sum_{w \in \mathcal{P}_t} w < \infty$. It remains to be proven, that the norm of the Hessian matrix is bounded by the Lipschitz constant L or the above Eqn. 5.49 holds true.

Without loss of generality, we can say that the space of gradients and parameters belongs to an $\mathbb{D}$-dimensional real vector space. It can be proved empirically that with every different initialization $\mathcal{P}_0$, the endpoint $\mathcal{P}_T$ differs. Since the parameter space is a real space, we can say that the sequence $\{\mathcal{P}_0^{(1)}, \mathcal{P}_1^{(1)}, \ldots, \mathcal{P}_T^{(1)}\}$ obtained with seed s_1 is disjoint from the sequence $\{\mathcal{P}_0^{(2)}, \mathcal{P}_1^{(2)}, \ldots, \mathcal{P}_T^{(2)}\}$ obtained with seed s_2, where $s_1 \neq s_2$. Thus, we can say that the vector space of parameters $\mathcal{P}$ is a Hausdorff Topological Vector Space with the canonical metric $d = \|\cdot\|$ of a normed space $(X, \|\cdot\|)$.

Since any two sequences of parameters on $\mathbb{R}^\mathbb{D}$ are disjoint, the gradient space associated with the sequences will also be disjoint on $\mathbb{R}^\mathbb{D}$. Hence, the gradient space defined on $\mathbb{R}^\mathbb{D}$ is also a Hausdorff Topological Vector Space with the canonical metric $d = \|\cdot\|$ of a normed space $(X, \|\cdot\|)$.

Since the parameter space $\mathcal{P}$ and the gradient space $\mathcal{G}$ are both Hausdorff spaces, the sequence converges to a point in the same respective space. In other words, the sequence $(\mathcal{P})_{t=1}^{\infty}$ and $(\mathcal{G})_{t=1}^{\infty}$ converges to some $\mathcal{P}_{t=\infty} \in \mathcal{P}$ and $\mathcal{G}_{t=\infty} \in \mathcal{G}$, respectively. The aforementioned statement implies that the parameter space $\mathcal{P}$ and the gradient space $\mathcal{G}$ are Banach spaces.

We can also view the above statement in another way. Since the composite function $(\mathcal{L}_{v3} \circ g_\psi \circ f_\theta)(x)$ belongs to the class of C^∞ functions, then under the constraint that the inputs to the Linear layers $\psi^{(1)}$ and $\psi^{(2)}$ are finite, i.e., $\sum_d h_{\theta n}^{(d)} < \infty$, and $\sum_{w \in \mathcal{P}} w < \infty$, the gradient values obtained using Eqn. 5.38 is also finite. Thus the change in consecutive values of a sequence in $\mathcal{P}$, i.e, $\|\mathcal{P}_{t+1} - \mathcal{P}_t\| < r_\mathcal{P}$, where $r_\mathcal{P} > 0$ and $r_\mathcal{P} \in \mathbb{R}$. Therefore, the sequence of parameters $(\mathcal{P}_t)_{t=1}^{\infty}$ can be called to be Cauchy in $(\mathcal{P}, \|\cdot\|)$. Thus, we can conclude that the parameter space $\mathcal{P}$ is a Banach space.

Similarly, since the gradients are finite, we can say that, $\|\nabla\mathcal{L}_{v3} \circ (g \circ f)_{\mathcal{P}_{t+1}}(x) - \nabla\mathcal{L}_{v3} \circ (g \circ f)_{\mathcal{P}_t}(x)\| < r_g$, where $r_g > 0$ and $r_g \in \mathbb{R}$. Therefore, denoting $\nabla\mathcal{L}_{v3} \circ (g \circ f)_{\mathcal{P}_{t+1}}(x)$ by $\mathcal{G}_{t+1}$ and $\nabla\mathcal{L}_{v3} \circ (g \circ f)_{\mathcal{P}_t}(x)$ by $\mathcal{G}_t$ we can proceed as,

$$\|\mathcal{G}_{t+1} - \mathcal{G}_t\| < r_g^{(t)} \leq L_g \cdot \|\nabla\mathcal{L}_{v3} \circ (g \circ f)_{\mathcal{P}_t}(x)\| \leq L_g^{'} \|\mathcal{P}_{t+1} - \mathcal{P}_t\| \tag{5.51}$$

where $\mathcal{G}$ is a point in the gradient space $\mathbb{G}$ and $\mathcal{G}_t \in \mathbb{G}$ denotes the state of the gradient at time step t.

Thus, the composite function approximated by $\mathcal{L}_{v3} \circ g_\psi \circ f_\theta : \chi \to \mathbb{R}$ has a $L_g^{'}$-Lipschitz continuous gradient, under the constraint that $\sum_d h_{\theta n}^{(d)} < \infty$, and $\sum_{w \in \mathcal{P}} w < \infty$. In other words, the above discussion indicates that the aforementioned function has locally Lipschitz continuous gradient when the weights are initialized with weights from a normal distribution. A different proof to arrive at the same conclusion is also provided in Lemma 2.3 in the concurrent work Patel and Berahas (2024).

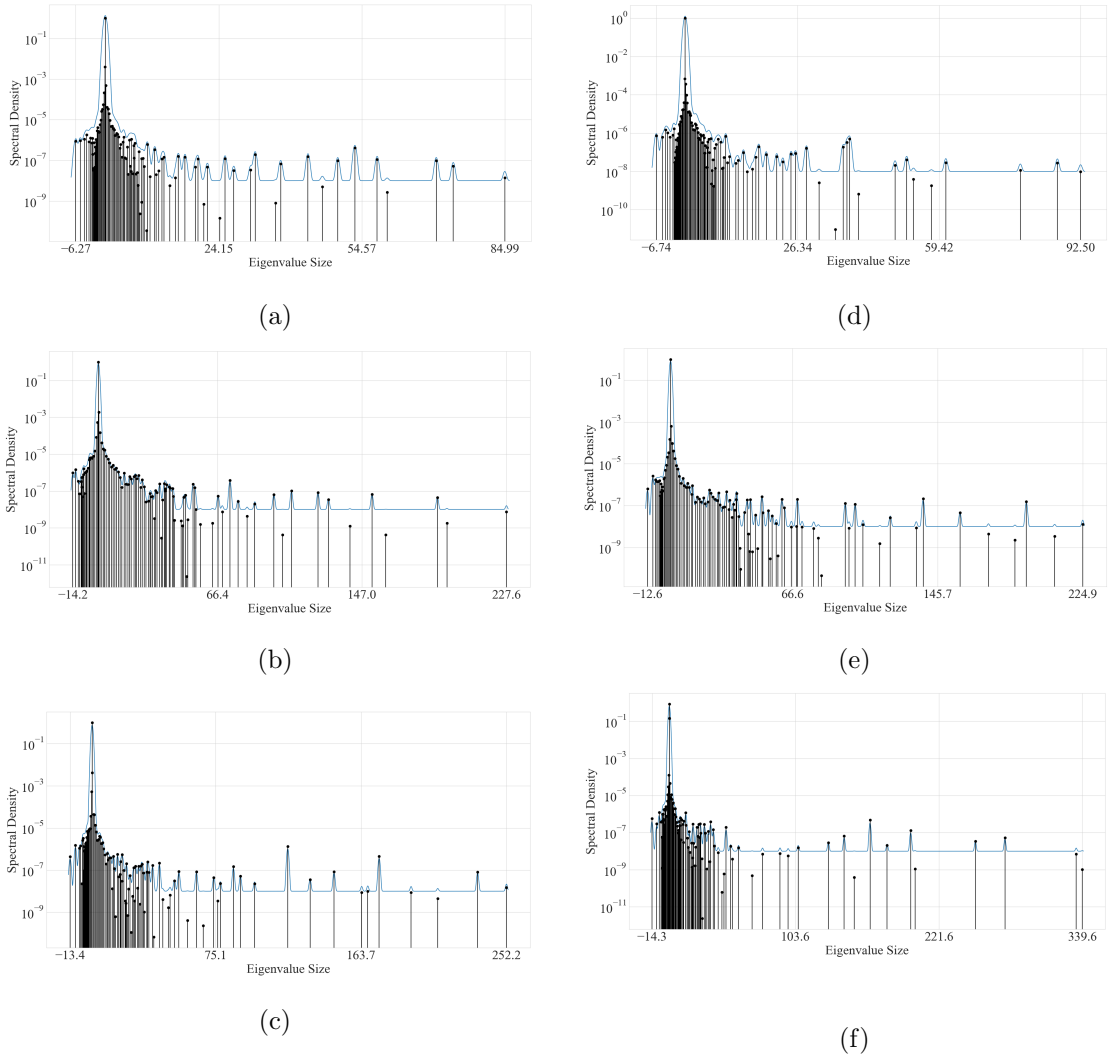


FIGURE 5.3: Plot of eigenvalues of parameters of ResNet18, obtained after 10 epochs of pre-training on CIFAR10 (a, b, c) and CIFAR100 (d, e, f) datasets with different SSL frameworks, namely, SimCLR (a, d), DCL (b, e) and MIOv3 (c, f).

Do contrastive SSL methods converge? As the learning rate decreases, the conditions become more favourable for the descent into a convex valley in the loss landscape. However, decreasing the learning rate deters the optimizer from proceeding with the same ease on flat plateaus or at inflection points to escape local minima. To ensure convergence along the steepest eigendirection, it is necessary to have a learning rate $\eta \leq \frac{1}{L} = \frac{1}{\lambda_{max}}$ (Bottou et al., 2018), where L is the Lipschitz constant.

Following Karimi et al. (2016), rewriting the Polyak-Lojasiewicz Inequality (See Appendix A for detailed discussion) in terms of loss function $\mathcal{L}$, for $\mu > 0$ the linear convergence rate is given by

$$\mathcal{L}(w_t) - \mathcal{L}^* \leq (1 - \frac{\mu}{L})^t (\mathcal{L}(w_0) - \mathcal{L}^*) \tag{5.52}$$

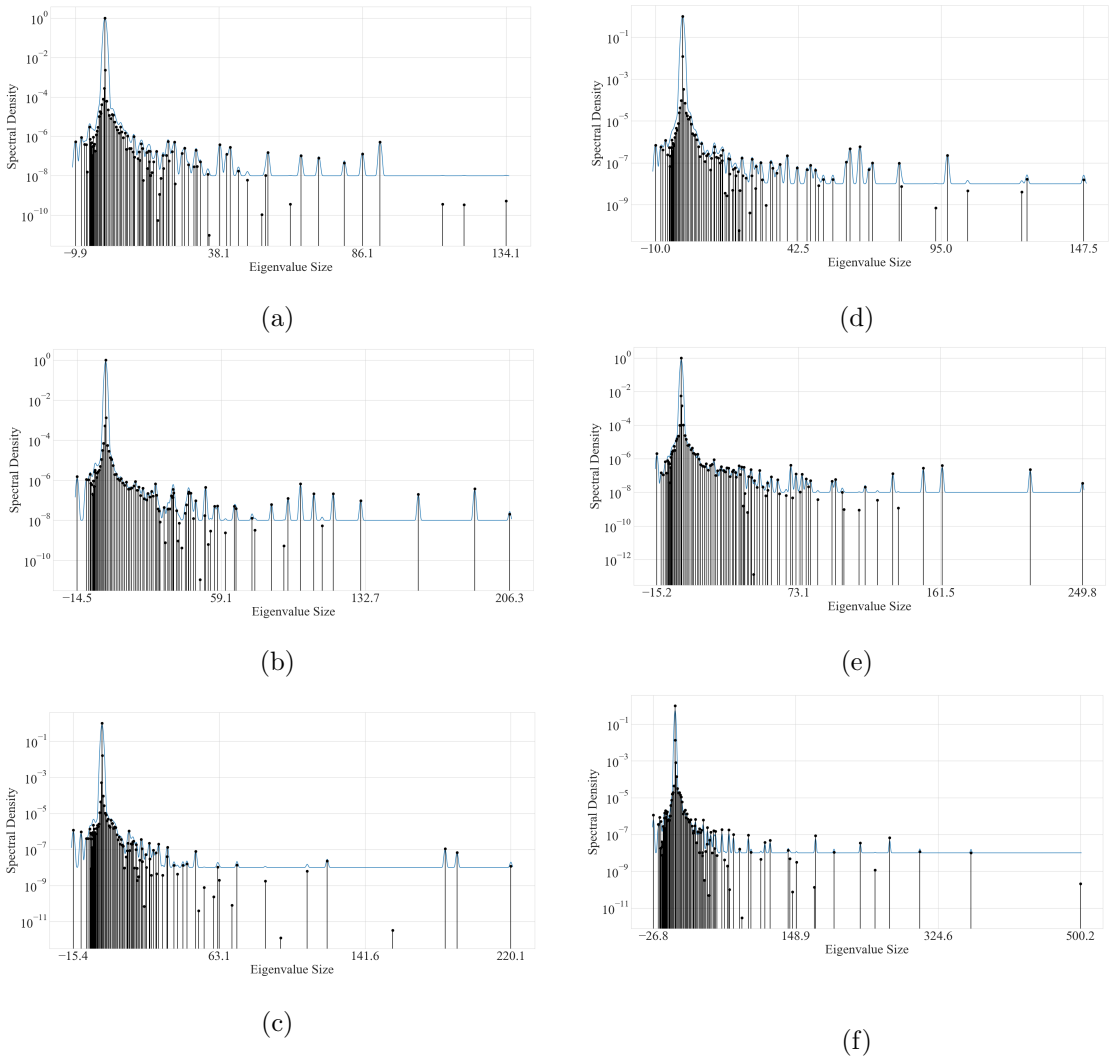


FIGURE 5.4: Plot of eigenvalues of parameters of ResNet18, obtained after 100 epochs of pre-training on CIFAR10 (a, b, c) and CIFAR100 (d, e, f) datasets with different SSL frameworks, namely, SimCLR (a, d), DCL (b, e) and MIOv3 (c, f)..

where w_t, w_0 are the parameter state at the t^{th} and 0^{th} step. For gradient descent algorithm or minimization problems, $L > 0$ always, at the minimum, and $\mu < L$.

In Lee et al. (2016), it is stated that a twice differentiable continuous function which is initialized randomly converges to a local minimum *almost surely*. Thus, given a *step size small enough*, we can derive the convergence rate in the local neighbourhood of the initialization point for the function approximated by the deep neural network. Therefore, as mentioned in Karimi et al. (2016), we can analyze the convergence phase in terms of locally satisfying the PL inequality.

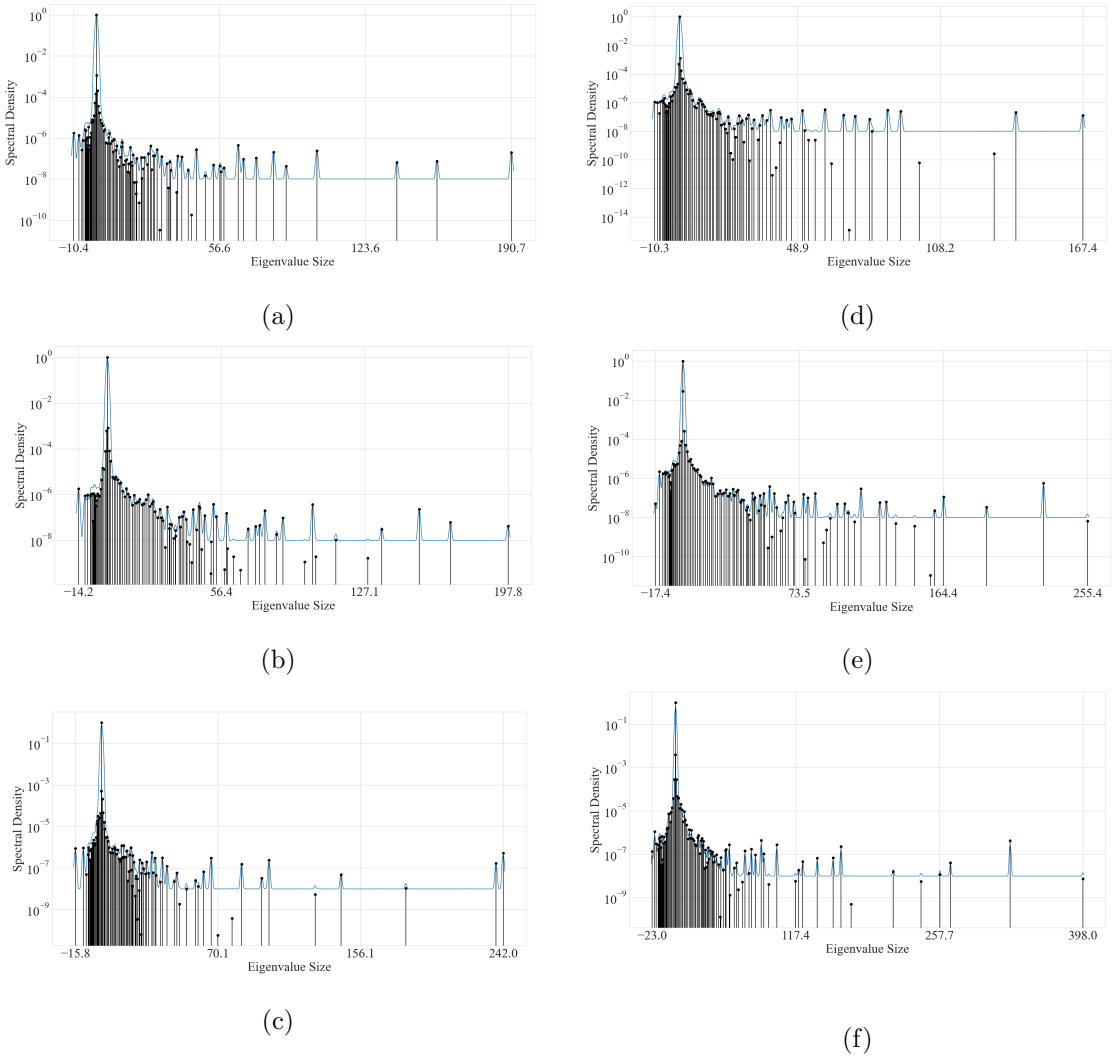


FIGURE 5.5: Plot of eigenvalues of parameters of ResNet18, obtained after 200 epochs of pre-training on CIFAR10 (a, b, c) and CIFAR100 (d, e, f) datasets with different SSL frameworks, namely, SimCLR (a, d), DCL (b, e) and MIOv3 (c, f)..

If we consider convergence along each eigendirection, we can calculate the expected convergence rate across the whole loss landscape. Considering a single eigendirection corresponding to maximum eigenvalue λ_i, the above equation reduces to,

$$\mathcal{L}(w_t^i) - \mathcal{L}^*(w^i) \leq (1 - \frac{\mu_i}{\lambda_i})^t (\mathcal{L}(w_0^i) - \mathcal{L}^*(w^i)) \tag{5.53}$$

subject to the satisfiability of

$$\frac{1}{2}\|\nabla \mathcal{L}^i(w)\|^2 \geq \mu_i(\mathcal{L}^i(w) - \mathcal{L}^*) \text{ for } \mu_i > 0 \tag{5.54}$$

We take an expectation over all the eigendirections to calculate a proxy for the linear convergence rate.

$$\begin{aligned}
\mathbb{E}_i\left[\mathcal{L}(w_t^i) - \mathcal{L}^*(w^i)\right] &\leq \mathbb{E}_i\left[(1 - \frac{\mu_i}{\lambda_i})^t(\mathcal{L}(w_0^i) - \mathcal{L}^*(w^i))\right] \\
&\leq \mathbb{E}_i\left[(1 - \frac{\mu_i}{\lambda_{max}})^t(\mathcal{L}(w_0^i) - \mathcal{L}^*(w^i))\right] \\
&\leq \mathbb{E}_i\left[(1 - \frac{\mu_i}{L})^t(\mathcal{L}(w_0^i) - \mathcal{L}^*(w^i))\right] \\
&\leq \mathbb{E}_i\left[(1 - \frac{\mu_{min}}{L})^t(\mathcal{L}(w_0^i) - \mathcal{L}^*(w^i))\right] \\
&\leq \mathbb{E}_i\left[c^t(\mathcal{L}(w_0^i) - \mathcal{L}^*(w^i))\right] \leq \delta' < \infty
\end{aligned} \tag{5.55}$$

where $c^t = (1 - \frac{\mu_{min}}{L})^t \to 0$ if $t \to \infty$, as $\mu_{min} \to 0^+$, and $\mu_i > 0 \; \forall \; i$.

From Equation (5.55), we can see that the convergence rate becomes infinitesimal for a large value of t, that is, for a long training process. However, we will look at Eqn. A.11 of the Appendix A, where we derive an expression of the expected gradient norm, as

$$\sum_{t=1}^{T}\left(\eta_t - \frac{\eta_t^2 L}{2}\right)\mathbb{E}_t\left[\|\nabla\mathcal{L}^i(w)\|_2^2\right] \leq \mathcal{L}(w_t) - \mathcal{L}^* + \frac{\sigma^2 L}{2}\sum_{k=1}^{T}\eta_t^2 \tag{5.56}$$

where we have assumed that the variance of the stochastic gradient is bounded above by σ^2.

Putting the expression for the proxy of the convergence rate in place of $\mathcal{L}(w_i) - \mathcal{L}^*$, we get,

$$\sum_{t=1}^{T}\left(\eta_t - \frac{\eta_t^2 L}{2}\right)\mathbb{E}_i\left[\|\nabla\mathcal{L}^i(w)\|_2^2\right] \leq \mathbb{E}_i\left[\mathcal{L}(w_t^i) - \mathcal{L}^*(w^i)\right] + \frac{\sigma^2 L}{2}\sum_{t=1}^{T}\eta_t^2 < \infty \tag{5.57}$$

where, η_t varies with time as $\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\frac{t}{T}\pi\right)\right)$ with $\eta_{min} = 0$, and T is the total number of steps. From the expression of η_t, we have,

$$\sum_{t=1}^{\infty}\eta_t \to \infty \tag{5.58}$$

$$\sum_{t=1}^{\infty}\eta_t^2 < \infty \tag{5.59}$$

From Eqns. (5.55) and (5.57), we calculate a proxy for the linear convergence rate and deduce that only longer training results in convergence provided Eqn. (5.52) is satisfied.

Furthermore, the fact that limited pre-training does not lead to convergence in self-supervised learning can be shown through a visualization of the eigenvalues of the parameters. In Fig. 5.3, 5.4 and 5.5, we plot the eigenvalues of the parameters for different

self-supervised frameworks after pre-training for 10, 100 and 200 epochs, respectively on both the CIFAR10 and CIFAR100 datasets, and we see that a small proportion of eigenvalues are negative. Although the mathematical results of our work are also supported by the findings presented in Pascanu et al. (2014) and Lee et al. (2016), the plots in Fig. 5.5 show otherwise. Consequently, this indicates that the parameter state at the end of pretraining does not converge to local minima along some eigendirections. Although according to the findings in Lee et al. (2016), the function should have converged to a local minimizer along every eigendirection *almost surely*. This phenomenon points towards the role of step size causing optimization along some eigendirections to get stuck at a saddle point or failing to escape out of the same, as discussed previously.

For computing the eigenvalues of the Hessian of the model parameters, we use the Deep Curvature suite (Granziol et al., 2019), which uses the Lanczos iteration method to compute the 100 eigenvalues which approximate the entire eigenspectrum. The 100 eigenvalues computed using this method are not the highest or lowest eigenvalues. Rather, these 100 eigenvalues cover the whole eigenspectrum and are thus very useful in developing a proper idea about the parameter state and how far it is from convergence.

5.5.4 Ablation Studies

5.5.4.1 Effect of Temperature

In this section, we study the effect of temperature on the proposed loss and also analyse the behaviour of MIOv1, MIOv2 and MIOv3 with varying temperature hyper-parameter values. We also notice the performance of MIOv1, MIOv2, and MIOv3 on different small-scale datasets (CIFAR-10, CIFAR-100, STL-10) and ImageNet-100 for different temperature values. As discussed in Sec. 5.4.2 and 5.4.3, we observe that the performance of MIOv2 is better than MIOv1 and MIOv3 at lower temperatures. But at higher temperatures, MIOv1 performs better than MIOv2. Furthermore, the drop in performance of MIOv3 at low temperature values is primarily due to increased repulsion between samples in false negative pairs. It is noteworthy that MIOv3 achieves the best performance at temperature 0.2 and also outperforms contemporary state-of-the-art SSL frameworks. Temperature hyper-parameter values higher than 0.5 are not used in our experiments, as the performance starts to drop at temperature 0.5.

5.5.4.2 Effect of Training Duration

In this subsection, we study the effect of training duration on the performance of the contemporary and proposed contrastive SSL frameworks. We observe from the empirical results presented in Table 5.9, that the proposed framework outperforms SimCLR and SimCLR+DCL when pre-trained for 200 and 1000 epochs.

TABLE 5.8: Variation of performance of MIOv1, MIOv2, MIOv3 for different temperature values, supporting the effect of $\mathcal{R}_{pp}$ and $\mathcal{R}_{nn}$ as described in Sec. 5.4.2 and 5.4.3.

Dataset	MIOvx	Temperature			
		0.07	0.1	0.2	0.5
CIFAR-10	v1	72.55	76.19	80.72	82.80
	v2	82.8	**82.87**	81.12	79.34
	v3	38.4	81.55	**86.36**	**85.09**
CIFAR-100	v1	36.3	42.9	50.7	49.7
	v2	47.1	46.2	43.5	35.8
	v3	33.6	**50.1**	**58.2**	**53.5**
STL-10	v1	62.8	67.3	72.42	71.78
	v2	71.33	**71.3**	70.2	65.0
	v3	32.1	71.0	**75.83**	**73.8**
IN-100	v1	-	65.04	71.98	74.02
	v2	-	75.24	73.80	70.98
	v3	-	**76.16**	**78.4**	**77.20**

TABLE 5.9: Comparison of SimCLR, DCL and MIOv3 on CIFAR10 and CIFAR100 datasets pre-trained for 200 and 1000 epochs.

Epochs	Dataset	SimCLR	DCL	MIOv3
200	CIFAR10	81.23	84.43	**86.36**
1000		89.5	88.16	**89.89**
200	CIFAR100	52.99	54.24	**58.18**
1000		60.5	61.03	**62.99**

5.5.4.3 Effect of Batch Size

In this subsection, we study the effect of varying batch sizes on the performance of the proposed contrastive SSL frameworks. We observe from the empirical results presented in Table 5.10, that the proposed framework performs best with batch size 128 on CIFAR10 and 256 on CIFAR100. However, the result with batch size 128 is presented in Table 5.2 for a fair comparison.

TABLE 5.10: Ablation of 200-NN Top-1 accuracy on CIFAR-10 and CIFAR-100 datasets for batch sizes of 64, 128, 256, and 512.

Dataset	Method	Batch size			
		64	128	256	512
CIFAR10	MIOv3	**85.9**	**86.28**	**86.19**	**85.9**
	DCL	84.32	84.43	84.4	83.86
	SimCLR	81.12	81.23	81.4	81.3
CIFAR100	MIOv3	**55.79**	**56.97**	**57.51**	**56.72**
	DCL	54.23	54.24	56.2	55.8
	SimCLR	51.66	52.99	53.6	53.69

TABLE 5.11: 200-NN accuracy of MIOv3, SimCLR, SimCLR+DCL, BYOL frameworks using 2 different models with decreasing number of parameters on CIFAR-10 and CIFAR-100 datasets obtained after 500 and 200 epochs of pre-training, respectively with a batch size of 128.

	ResNet-18	ResNet-9
# Basic Blocks	[2,2,2,2]	[1,1,1,1]
Base Channels	64	64
# Params	$\sim$ 11M	$\sim$ 5M
Model	Top-1 Accuracy (%)	
	CIFAR-10	
MIOv3	**89.00**	**84.75**
SimCLR	84.97	80.18
SimCLR+DCL	86.4	82.82
BYOL	90.13	84.56
	CIFAR-100	
MIOv3	**58.18**	**53.98**
SimCLR	52.99	48.19
SimCLR+DCL	54.24	51.21
BYOL	54.02	50.98

5.5.4.4 Effect of Number of Parameters

In this ablation study, we mainly discuss the effect of the number of parameters on performance. Neural networks are in general over-parameterized. In this section, we intend to conduct an experimental study to determine the efficiency of parameter utilization in SSL. With a decrease in the number of parameters, the performance will surely drop. The performance of any particular framework with decreased parameters implies how much of the total number of parameters is being utilized by the framework for representation learning.

In Table 5.11, we have presented the 200-NN accuracy of MIOv3, SimCLR, SimCLR+DCL, and BYOL on the CIFAR10 and CIFAR100 datasets. The configuration of the base encoder is also mentioned in the table, along with the number of parameters. Intuitively, a decrease in the number of parameters will eliminate some useful parameters. However, the ability of the respective frameworks to utilize the previously redundant parameters can be observed from the performance as given in Table 5.11.

We see that the performance of all frameworks decreases with the decrease in the number of parameters. However, it is worth noting that, under the effect of decreasing the number of parameters, our proposed MIOv3 framework outperforms all self-supervised learning frameworks on the CIFAR dataset.

5.6 Conclusion

The main contribution of this chapter is twofold: (a) we proposed a novel binary contrastive loss function (MIOv3) that optimizes the mutual information between samples in positive and negative pairs, and (b) present an analysis of the convergence of contrastive SSL frameworks. Initially, we started from the base version MIOv1 and justified its formulation with reference to the noise contrastive estimation principle. We then modified it gradually by considering the uniformity and alignment metrics as references and finally obtained MIOv3 with superior performance. Through mathematical calculation, we provide a lower bound of the base loss function MIOv3, which is the difference between the mutual information of the samples in the negative and positive pairs, and also justifies our motivation of optimizing mutual information between samples in both positive and negative pairs. The proposed framework yields better results on both small-scale and large-scale datasets than the state-of-the-art instance discrimination-based contrastive, negative-free contrastive and non-contrastive frameworks. The results obtained by MIOv3 demonstrate that the proposed binary contrastive learning framework is better at optimizing the mutual information between the different types of pairs than most contemporary SSL frameworks. We also study the effect of temperature hyper-parameter, training duration, batch size, and decrease in model parameters on the downstream performance and notice that our proposed framework outperforms the contemporary frameworks in all scenarios.

From the eigenspectrum analysis we also observed how the optimization process proceeds on the parameter space in self-supervised learning. The primary idea was to investigate whether the SSL frameworks achieve convergence in the parameter space. To derive the convergence criterion for the SSL frameworks, we start with deriving the gradient of MIOv3 with respect to the parameters and then derive the hessian of MIOv3. The hessian aids in figuring out the conditions of Lipschitz continuity in the parameter space. Finally, by applying the Polyak-Lojasiewicz inequality in the local neighbourhood of the parameter state without violating the invexity assumption, we show both mathematically and empirically that under a longer duration of the training, SSL frameworks converge to strict saddle points in the loss landscape.

The development of MIOv3 from MIOv1 via MIOv2 was aided by the optimization of uniformity and alignment of the samples in the feature space. However, the uniformity and alignment metrics are heavily dependent on the temperature hyperparameter (Wang and Isola, 2020; Huang et al., 2023a), and influence the performance of the contrastive learning frameworks as evident from SimCLR (Chen et al., 2020a), MoCo (He et al., 2020), as well as from the results presented in this chapter. Hence, in the next chapter, we will try to figure out the effect of temperature hyper-parameter on the distribution of samples in the feature space and develop a cosine similarity-dependent temperature scaling function which improves the uniformity and alignment of the samples.

Chapter 6

Dynamic Temperature Hyper-Parameter Scaling in Self-Supervised Contrastive Learning

6.1 Introduction

Self-supervised learning (SSL) has become a cornerstone in machine learning due to its ability to learn high-quality representations from large-scale unlabeled data. Initial approaches in SSL involved designing a suitable pretext task, such as solving jigsaw puzzles (Noroozi and Favaro, 2016), image inpainting (Pathak et al., 2016), colourization (Zhang et al., 2016), etc. However, the problem with these methods was that there exists a significant difference between the nature of the pretext task and the desired downstream task (classification/segmentation), resulting in quite a considerable performance gap between the SSL frameworks and their supervised counterparts. Contrastive learning framework (He et al., 2020; Chen et al., 2020a; Yeh et al., 2022), wherein the model learns an embedding space such that the features of augmented versions of the same sample lie close to each other and pushes the embeddings of dissimilar samples farther apart, have revolutionized self-supervised learning and greatly increased the applicability of SSL to different tasks. Empirically, contrastive learning-based algorithms have been found to perform better at downstream tasks than the former SSL methods and at par with their supervised counterparts.

The most commonly used loss function for self-supervised contrastive learning (SSCL) is the InfoNCE loss, first introduced in the self-supervised context in CPC (van den Oord et al., 2018) and subsequently adopted in several works such as MoCo (He et al., 2020), SimCLR (Chen et al., 2020a), etc. The temperature hyper-parameter of the InfoNCE function plays a significant role in controlling the quality of representations in the pre-training phase. However, there is a lack of proper study on the same. In the recent past, there have been a few pieces of work that have focused on the temperature hyper-parameter in the

InfoNCE loss function. In Zhang et al. (2021b), the authors present a temperature hyper-parameter as a function of the input representations thereby incorporating uncertainty in the form of temperature. Wang and Liu (2021) explores the hardness-aware property of contrastive loss and the role of temperature hyper-parameter in it by measuring the uniformity and tolerance of representations. On the other hand, MACL Huang et al. (2023a) assumed the temperature hyperparameter as the function of alignment to address the uniformity-tolerance dilemma that arises as a result of optimizing the InfoNCE loss in a self-supervised setting. Motivated by the study shown in Wang and Liu (2021), the authors in Kukleva et al. (2023) proposed a continuous task switching between instance discrimination and an implicit group-wise discrimination by using simple cosine scheduling. However, the framework in Kukleva et al. (2023) requires a longer training duration than conventional SSL frameworks. In Qiu et al. (2023), the authors attempt to implement distributionally robust optimization for individual temperature individualization, that is, it uses a temperature hyper-parameter τ_i corresponding to each anchor sample x_i, and is updated at each iteration. This no longer keeps the temperature as a hyper-parameter and converts it into another optimizable parameter.

In this chapter, however, we argue that there is more significance to this temperature hyper-parameter than initially apparent. Our theoretical analyses bring forth an intuitive yet vital aspect related to the presence of constructively false negative yet inherently positive pairs – samples that do not originate from the same instance yet show a high degree of semantic representational similarity as they belong to the same underlying class. As the main objective of the contrastive loss function is to maximize the similarity of the different augmentations of the same instance while minimizing the same for different instances, the aforementioned constructively false negative pairs are repelled away. This action implies that semantic information is not an integral part of existing InfoNCE loss. Pushing the samples in semantically similar pairs away creates an adverse effect on representation learning. Large penalties on these samples along with true negative samples may increase the uniformity, but it adversely affects the alignment of the local structure constituted by samples with similar semantic information. Hence the uniformity-tolerance (alignment) dilemma arises as addressed in Wang and Liu (2021). We also intend to dynamically scale the temperature hyper-parameter as a function of the cosine similarity to effectively control the repelling effect in these false negative pairs. We theorize that scaling the temperature dynamically will prevent disruption of the local and global structures of the feature space and improve representation learning. To this end, we systematically study the role of temperature hyper-parameter and its effect on local and global semantic structures in the feature space during optimization of the InfoNCE loss, both intuitively and theoretically, to establish the motivation for our proposed method. With the established groundwork, we propose a temperature-scaled contrastive learning framework (DySTreSS) that dynamically modulates the temperature hyper-parameter based on local and global structures. Furthermore, we show the effectiveness of our approach by conducting experimentation across several benchmark vision datasets, with empirical results showing that our method outperforms better than several state-of-the-art SSL algorithms in the literature. A part of the results reported in this chapter are also presented in Manna et al. (2024).

The rest of the chapter is organized as follows: Sec. 6.2 discusses the preliminaries required for understanding the theoretical development discussed in the following sections.

Sec. 6.3 presents the motivation driving the development of the proposed dynamic temperature scaling function in this chapter. Sec. 6.4 deals with the mathematical formulation and also introduces the proposed framework. Next, we present the experimental details, report comparative results with detailed analyses, and include ablation studies in Sec. 6.5. Finally, we conclude our work in Sec. 6.6.

6.2 Preliminaries

InfoNCE Loss In self-supervised contrastive learning (van den Oord et al., 2018; Chen et al., 2020a; He et al., 2020), the InfoNCE loss is given by Eqn. 6.1.

$$\mathcal{L} = \sum_i \mathcal{L}_i = -\sum_i \ln\left(p_{ii+}\right) = -\sum_i \ln\left(\frac{exp(\frac{c_{ii+}}{\tau})}{\sum_j exp(\frac{c_{ij}}{\tau})}\right) \tag{6.1}$$

where $ii+$ denotes a true positive pair and c_{ij} is the cosine similarity between the latent vectors of the samples x_i and x_j.

Nomenclature of Sample Pairs In self-supervised contrastive learning frameworks like SimCLR (Chen et al., 2020a), MoCo (He et al., 2020), etc., it is assumed that each sample belongs to a different class. Hence, when any two samples are paired, it may result in the pairing of two samples that may belong to the same class, resulting in the formation of false negative (FN) pairs. The similarity between these samples in these types of pairs can yield high cosine similarity values. On the contrary, pairs consisting of two samples belonging to two different classes comprise true negative (TN) pairs. However, depending on the mapping of the corresponding features to the feature space, true negative pairs can also have high cosine similarity between the constituent samples and can be treated as hard true negative pairs. False negative pairs by construction can also act as hard false negative pairs if they are mapped far from each other resulting in low cosine similarity. True positive (TP) pairs are simply constituted of samples obtained by two random augmentations of a sample in the dataset.

6.3 Motivation

6.3.1 Role of Temperature in Contrastive Learning

InfoNCE loss concentrates on optimization by penalizing the hard negative pairs according to their hardness (Wang and Liu, 2021). The gradient of $\mathcal{L}_i$ w.r.t. c_{ii} and c_{ij} is given by Eqn. 6.2 and 6.3. A simple relative weightage is defined in Wang and Liu (2021) and as given in Eqn. 6.4.

$$\frac{\partial \mathcal{L}_i}{\partial c_{ii+}} = -\frac{1}{\tau}\sum_{k\neq i} p_{ik} = -\frac{1}{\tau}(1 - p_{ii}) \tag{6.2}$$

$$\frac{\partial \mathcal{L}_i}{\partial c_{ij}} = \frac{1}{\tau} p_{ij} \tag{6.3}$$

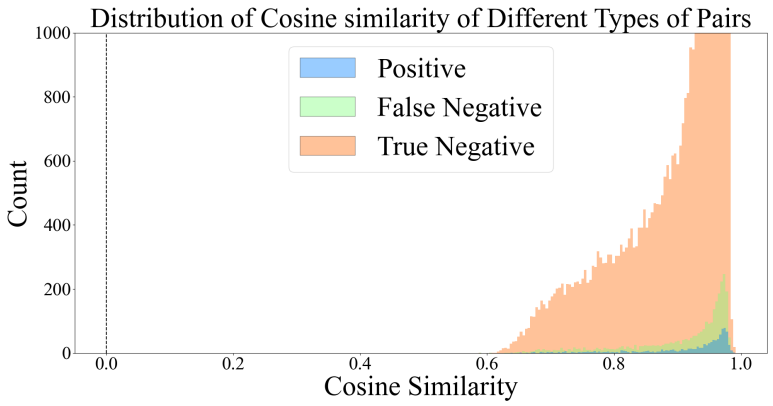
$$r(c_{ij}) = \frac{|\frac{\partial \mathcal{L}_i}{\partial c_{ij}}|}{|\frac{\partial \mathcal{L}_i}{\partial c_{ii+}}|} = \frac{\exp(\frac{c_{ij}}{\tau})}{\sum_{k \neq i} \exp(\frac{c_{ik}}{\tau})} \tag{6.4}$$

Therefore, the role of temperature τ in contrastive loss is to control the relative weightage of the hard negative samples. The entropy of the penalty distribution is a monotonically increasing function w.r.t. τ. A low temperature τ also results in sharp $r(c_{ij})$ in a high similarity region, giving higher penalties to samples close to the anchor sample x_i. The penalty distribution is more uniform at higher temperatures, which gives the negative samples the same magnitude of penalties. As the penalty distribution follows a Boltzmann distribution, the effective penalty grows exponentially as the temperature decreases with the cosine similarity value. Hence, for any anchor sample x_i only the nearest few samples are penalized. Thus, hard negative (HN) pairs with either similar (false negative) or dissimilar (true negative) semantic features and high cosine similarity are penalized more. However, penalizing FN pairs disturbs the structure of the clusters in the feature space, by generating large gradients for the closely located false negative samples. On the other hand, high temperature values tend to create closely located clusters in the feature space by smoothing out the gradients given by Eqn. 6.3, between closely located false negative pairs. Thus, lower temperatures tend to generate more uniformly distributed samples in the feature space. Hence, it is evident that the temperature hyper-parameter acts as the control knob for the uniformity. Alternatively, the entropy of $r(c_{ij})$ is monotonically increasing the temperature hyper-parameter (Wang and Liu, 2021). Thus, as the temperature decreases, the entropy of $r(c_{ij})$ is also decreasing, indicating a uniform distribution of samples in the feature space.

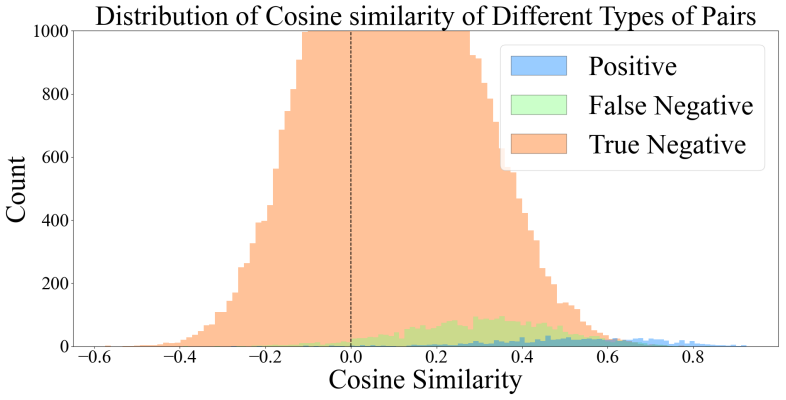
6.3.2 Effect of Temperature on Local and Global Structures

Let us assume that for a given sample x, we have an encoder $f_\theta : x \to z \in \mathbb{R}^D$, where z is a mapped image and θ is the encoder parameters. Under any valid distance measure $\mathfrak{U}$ on the manifold $\mathfrak{M}$ of z, in an optimal scenario, if convergence is achieved in a self-supervised pre-training stage, two mapped images z_i and z_j, where $i \neq j$, from same class C, will have minimum possible distance. In our work, the term 'local structure' of any sample x_j refers to the arrangement of the other samples in the close local neighborhood of that sample and can be denoted by the samples included in a closed ball of radius r_j (> 0) around that sample x_j. Likewise, the term 'global structure' takes the arrangement of all the samples in the feature space into account.

As already stated in the previous subsection , decreasing the temperature tends to penalize the hard negative pairs more. This is because hard negative pairs tend to have high cosine similarity (say, c_{ij}). With small temperature, the quantity $\frac{c_{ij}}{\tau}$ is further amplified, consequently resulting in a larger penalty (from Eqn. 6.4). This causes samples constituting false negative pairs to drift apart. Consequently, the local structure consisting of samples of any particular class is disturbed. However, we would want the samples in false negative pairs to stay close, that is, the tolerance needs to be higher for better linear separability in



(a)



(b)

FIGURE 6.1: (a) Histogram of cosine similarities of true positive (TP), false negative (FN), and true negative (TN) pairs at random initialization, (b) Histogram of cosine similarities of TP, FN, and TN pairs after pre-training.

the feature space. This effect on false negative pairs gives rise to the "uniformity-tolerance dilemma".

The effect of temperature can be better understood if we take the gradient of the loss with respect to any latent vector z_j. Taking the expression of the derivative of the loss $\mathcal{L}$, given by Eqn. 6.1, with respect to z_j, we get Eqn. 6.5. When differentiating the infoNCE loss with respect to a feature vector z_i, we need to consider two things, (1) the term where z_i is the anchor, and (2) all the terms where z_i is not the anchor. This gives rise to the two terms in the first line of the differentiation in Eqn. 6.5.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial z_i} &= \left[-\frac{z_{i^+}}{\tau} + \frac{\frac{z_{i^+}}{\tau} \cdot e^{C'_{ii^+}} + \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{z_j}{\tau} \cdot e^{C'_{ij}}}{e^{C'_{ii^+}} + \sum_{\substack{j=1 \\ j\neq i}}^{N} e^{C'_{ij}}} \right] + \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{\frac{z_j}{\tau} \cdot e^{C'_{ji}}}{e^{C'_{jj^+}} + \sum_{\substack{k=1 \\ k\neq j}}^{N} e^{C'_{jk}}} \\ &= -\left[\frac{z_{i^+}}{\tau} \left(1 - p_{ii^+}\right) - \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{z_j}{\tau} \left(p_{i\Downarrow j} + p_{j\Downarrow i} \right) \right]\end{aligned} \tag{6.5}$$

where $C'_{ij} = \frac{c_{ij}}{\tau}$, denotes the cosine similarity between the feature vectors z_i and z_j, scaled by temperature τ, and (z_i, z_{i+}) forms the positive pair. The quantity $p_{i\Downarrow j}$ is the probability of the pair (x_i, x_j) being predicted as a positive pair with the sample x_i as the anchor. Hence, from the expression of the displacement vector $\frac{\partial \mathcal{L}}{\partial z_i}$, we can conclude that at a low-temperature value the sample z_i moves away from any sample z_j if they are mapped close to each other in the feature space. In other words, contrastive loss penalizes hard negative pairs. The effect of temperature reduction in different scenarios is discussed as follows.

Reducing Temperature for False Negatives: By gradient descent rule, the updated position of the sample x_i in the feature space can be denoted by $z_i^{t+1} = z_i^t - \frac{\partial \mathcal{L}}{\partial z_i^t}$. The value of cosine similarity c_{ij} between the two samples in a false negative pair can be positive or negative depending on where the samples are mapped in the feature space. Adjusting the temperature hyper-parameter τ allows us to control the contribution of the false negative pairs in the loss optimization process, by scaling the weights of the latent vectors z_j in Equation (6.5). For two closely placed false negative samples, the value of the term $p_{i\Downarrow j} + p_{j\Downarrow i}$ as shown in Equation (6.5) will be high. If the temperature is decreased, the contribution of the sample z_j in the gradient scales up further, resulting in the sample z_i drifting opposite to the direction of z_j. Conversely, if we take the derivative of $\mathcal{L}$ with respect to z_j similar to Equation (6.5), we will get a term involving z_i, which will enforce a similar effect on z_j. This results in the disruption of the local cluster structure in the feature space. This phenomenon increases uniformity but decrease in alignment or tolerance, resulting in degraded linear separability of classes in the feature (latent) space.

Reducing Temperature for Hard Negatives: For hard negative pairs, we can expect the two constituent samples to drift apart from each other if a low enough temperature is applied, as the cosine similarity between hard negative pair samples is generally high. Without a ground truth label, it is impossible to apply selective temperature moderation to all the pairs. If the temperature for all pairs whose cosine similarity is above a certain threshold (say, c_α) is decreased, then the closely spaced false negative pairs will also be affected, resulting in disruption of the local cluster structure.

Increasing Global Temperature: Increasing the temperature for all the samples has the opposite effect. As the temperature is increased, the drift in the false negative pairs is reduced, thereby helping in maintaining proper alignment. However, the uniformity may be affected as the repulsion between samples constituting true negative pairs including the hard true negatives, will also be reduced. Hence, increasing temperature causes an increase in alignment but affects uniformity. Similarly, if the temperature is increased only for pair samples with cosine similarity below the threshold c_α, then the true negative pair

samples are not efficiently repelled, resulting in a mapping with low linear separability of classes.

What if there were no False Negatives? An ideal scenario for contrastive learning would be the absence of any false negative pairs. In such a scenario, where all the negative pairs are true negative pairs (like in supervised contrastive learning (Khosla et al., 2020), we may make the mistake of assuming that we can safely decrease the temperature. Decreasing the temperature for true negative pairs will certainly improve performance up to a certain level, below which the performance degrades due to numerical instability (Khosla et al., 2020), as the gradients become too large. This degradation in performance is due to the disruption in both the local and global structure of the feature space, causing an increase in the uniformity of the sample features and a decrease in alignment. Disruption in local structure causes degradation of alignment in the feature space, whereas disruption in the global structure will cause an increase in uniformity (Wang and Isola, 2020; Wang and Liu, 2021).

6.3.3 Intuitive Tenets of Ideal Temperature Scaling Function

In SSCL, the boundary between true and false negatives cannot be distinguished with certainty. It is also not possible to infer the class labels. In the above subsections , we have discussed the effect of temperature on the feature space intuitively and can list some basic tenets which a proper temperature scaling function should follow. However, it is worth mentioning before proceeding further that we do not adhere to any such assumptions to formulate the proposed framework. The criteria are as follows: (1) *Local criteria*: A very low temperature in both the highly positive and negative cosine similarity region will disrupt the local structure, (2) *False negative criteria*: A very low temperature for false negative pairs, which we can assume to lie in the range $[c_{fn}, +1.0]$ can affect hard true negative and true positive pairs, where c_{fn} denotes a cosine similarity score, (3) *Global criteria*: A high temperature will affect the uniformity of the feature space and delay convergence.

6.4 Proposed Framework

6.4.1 Mathematical Formulation

Let us assume $\tau(\cdot)$ is the temperature function, which takes the cosine similarity of a pair as input and outputs a temperature value for the same. For the rest of this literature, we will consider $\tau_{ij} = \tau(c_{ij})$. Now, the temperature hyper-parameter τ being a function of the cosine similarity term c_{ij}, the expression for the derivative of the loss $\mathcal{L}$ would change, as given by Eqn. (6.6) and (6.7).

$$\frac{\partial \mathcal{L}_i}{\partial c_{ii+}} = -\frac{\partial}{\partial c_{ii+}} \left(\ln \frac{exp(\frac{c_{ii+}}{\tau_{ii+}})}{exp(\frac{c_{ii+}}{\tau_{ii+}}) + \sum_j exp(\frac{c_{ij}}{\tau_{ij}})} \right)$$

$$
\begin{aligned}
&= -\frac{1}{\left(\frac{exp(\frac{c_{ii+}}{\tau_{ii+}})}{exp(\frac{c_{ii+}}{\tau_{ii+}})+\sum_j exp(\frac{c_{ij}}{\tau_{ij}})}\right)} \frac{\partial}{\partial c_{ii+}} \left(\frac{exp(\frac{c_{ii+}}{\tau_{ii+}})}{exp(\frac{c_{ii+}}{\tau_{ii+}}) + \sum_j exp(\frac{c_{ij}}{\tau_{ij}})}\right) \\
&= -\frac{\mathfrak{C}}{exp(\frac{c_{ii+}}{\tau_{ii+}})} \left(\frac{\mathfrak{C}\frac{\partial}{\partial c_{ii+}} exp(\frac{c_{ii+}}{\tau_{ii+}}) - exp(\frac{c_{ii+}}{\tau_{ii+}})\frac{\partial}{\partial c_{ii+}}\mathfrak{C}}{\mathfrak{C}^2}\right) \\
&\text{where } \mathfrak{C} = exp(\frac{c_{ii+}}{\tau_{ii+}}) + \sum_j exp(\frac{c_{ij}}{\tau_{ij}}) \\
&= -\frac{\mathfrak{C}}{exp(\frac{c_{ii+}}{\tau_{ii+}})} \left(\frac{\mathfrak{C} \cdot exp(\frac{c_{ii+}}{\tau_{ii+}}) \cdot \frac{\partial \frac{c_{ii+}}{\tau_{ii+}}}{\partial c_{ii+}} - exp(\frac{c_{ii+}}{\tau_{ii+}}) \cdot exp(\frac{c_{ii+}}{\tau_{ii+}}) \cdot \frac{\partial \frac{c_{ii+}}{\tau_{ii+}}}{\partial c_{ii+}}}{\mathfrak{C}^2}\right) \\
&= -\frac{\mathfrak{C} \cdot \frac{\partial \frac{c_{ii+}}{\tau_{ii+}}}{\partial c_{ii+}} - exp(\frac{c_{ii+}}{\tau_{ii+}}) \cdot \frac{\partial \frac{c_{ii+}}{\tau_{ii+}}}{\partial c_{ii+}}}{\mathfrak{C}} = -\frac{\tau_{ii+} - c_{ii+}\frac{\partial \tau_{ii+}}{\partial c_{ii+}}}{\tau_{ii+}^2} \cdot (1 - p_{ii+})
\end{aligned}
\tag{6.6}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}_i}{\partial c_{ij}} &= -\frac{\partial}{\partial c_{ij}} \left(\ln \frac{exp(\frac{c_{ii+}}{\tau_{ii+}})}{exp(\frac{c_{ii+}}{\tau_{ii+}}) + \sum_j exp(\frac{c_{ij}}{\tau_{ii+}})}\right) \\
&= -\frac{1}{\left(\frac{exp(\frac{c_{ii+}}{\tau_{ii+}})}{exp(\frac{c_{ii+}}{\tau_{ii+}})+\sum_j exp(\frac{c_{ij}}{\tau_{ij}})}\right)} \frac{\partial}{\partial c_{ij}} \left(\frac{exp(\frac{c_{ii+}}{\tau_{ii+}})}{exp(\frac{c_{ii+}}{\tau_{ii+}}) + \sum_j exp(\frac{c_{ij}}{\tau_{ij}})}\right) \\
&= -\frac{\mathfrak{C}}{exp(\frac{c_{ii+}}{\tau_{ii+}})} \left(\frac{\mathfrak{C}\frac{\partial}{\partial c_{ij}} exp(\frac{c_{ii+}}{\tau_{ii+}}) - exp(\frac{c_{ii+}}{\tau_{ii+}})\frac{\partial}{\partial c_{ij}}\mathfrak{C}}{\mathfrak{C}^2}\right) \\
&\text{where } \mathfrak{C} = exp(\frac{c_{ii+}}{\tau_{ii+}}) + \sum_j exp(\frac{c_{ij}}{\tau_{ij}}) \\
&= -\frac{\mathfrak{C}}{exp(\frac{c_{ii+}}{\tau_{ii+}})} \left(\frac{-exp(\frac{c_{ii+}}{\tau_{ii+}}) \cdot exp(\frac{c_{ij}}{\tau_{ij}}) \frac{\partial \frac{c_{ij}}{\tau_{ij}}}{\partial c_{ij}}}{\mathfrak{C}^2}\right) = \frac{\tau_{ij} - c_{ij}\frac{\partial \tau_{ij}}{\partial c_{ij}}}{\tau_{ij}^2} \cdot p_{ij}
\end{aligned}
\tag{6.7}
$$

where c_{ii+} and c_{ij} denote the cosine similarity of positive and negative pairs, respectively.

We can intuitively say, without loss of generosity, that for negative pairs, the ideal behaviour is that the cosine similarity should decrease, and a decrease in the cosine similarity should cause a decrease in the loss too. Hence, we can write $\frac{\partial \mathcal{L}}{\partial c_{ij}} = \delta > 0$, where δ is a non-negative number. From Eqn. 6.7, we get,

$$
\frac{\partial \mathcal{L}}{\partial c_{ij}} = \frac{\tau_{ij} - c_{ij}\frac{\partial \tau_{ij}}{\partial c_{ij}}}{\tau_{ij}^2} p_{ij} = \delta \;\; \text{where } \delta < \epsilon \;\; \text{and } \delta, \epsilon > 0 \tag{6.8}
$$

Now, from the above Eqn. 6.8, we will try to verify the assumptions that we made about an ideal temperature function, and later we will show that these assumptions hold true

for our proposed temperature function by solving a differential equation arising from Eqn. 6.6 and 6.7. It is to be noted, that the assumptions about the slope of the temperature function do not influence the derivation in any way. Without loss of generosity, we can always assume $\tau_{ij} > 0$. As the temperature parameter cannot be negative or zero, the temperature value would be some positive constant. For $c_{ij} < 0$, to satisfy our criteria (1) and (2), we should have $\frac{\partial \tau_{ij}}{\partial c_{ij}}$ always less than some negative number. Hence, the slope of the temperature function is negative in the negative half of the cosine similarity vs. temperature plane. In the positive half of the cosine similarity vs. temperature plane, the slope of the temperature function is less than some positive number. However, a negative slope in the positive half would mean that temperature would decrease at high cosine similarity, again violating our criteria (1) and (2). A low temperature at high cosine similarity will affect the hard negative pairs and degrade the local structure. Taking into consideration the above two assumptions, we should adopt the temperature function such that the temperature does not violate criteria (3) at high cosine similarity values.

Now, we will derive our proposed temperature function in Proposition 1 simply from the basic gradient equations in Eqn. 6.6 and 6.7.

Proposition 1: *The temperature function should have a negative and positive slope in the negative and positive half of the cosine similarity vs. temperature plane, respectively.*

Proof: For negative pairs, the gradient of the InfoNCE loss with respect to c_{ij} will be non-negative, because, if loss decreases, then the cosine similarity of negative pairs should decrease. We assume that the value of this gradient is δ, as shown in Eqn. 6.9.

$$\frac{\partial \mathcal{L}}{\partial c_{ij}} = \frac{\tau_{ij} - c_{ij}\frac{\partial \tau_{ij}}{\partial c_{ij}}}{\tau_{ij}^2} p_{ij} = \delta \;\text{ where } \delta < \epsilon \;\text{ and } \delta, \epsilon > 0 \tag{6.9}$$

where ϵ is a small non-negative number. Expanding the Eqn. 6.9, we get,

$$\begin{aligned}
&\frac{\tau_{ij} - c_{ij}\frac{\partial \tau_{ij}}{\partial c_{ij}}}{\tau_{ij}^2} p_{ij} = \delta \\
\implies &\frac{\tau_{ij} - c_{ij}\frac{\partial \tau_{ij}}{\partial c_{ij}}}{\tau_{ij}^2} = \frac{\delta}{p_{ij}} \\
\implies &\tau_{ij} - c_{ij}\frac{\partial \tau_{ij}}{\partial c_{ij}} = \tau_{ij}^2 \frac{\delta}{p_{ij}} \\
\implies &\frac{\partial \tau_{ij}}{\partial c_{ij}} = \frac{1}{c_{ij}}\left[\tau_{ij} - \tau_{ij}^2 \frac{\delta}{p_{ij}}\right] \\
\implies &\frac{\partial \tau_{ij}}{\partial c_{ij}} = \frac{\tau_{ij}}{c_{ij}}\left[1 - \frac{\tau_{ij}\delta}{p_{ij}}\right]
\end{aligned} \tag{6.10}$$

We can assume that $\tau_{ij} > 0$ without loss of generality.

In self-supervised contrastive learning, the temperature should be high for false negatives to prevent too much repulsion. We have discussed the criteria and the motivation behind

our temperature function in Sec. 6.3.3 of the main manuscript. Also, the temperature should not be very small in the regions with highly negative cosine similarity. We assume that the number of false negatives decreases as we move towards the point $c_{ij} = 0.0$. Hence, for the vanilla case, we will consider two regions, (1) $c_{ij} > 0$ and (2) $c_{ij} \leq 0$.

Expanding the expression for p_{ij} in Eqn. 6.10, we get,

$$\begin{aligned}
&\frac{\partial \tau_{ij}}{\partial c_{ij}} = \frac{\tau_{ij}}{c_{ij}} \left[1 - \frac{\tau_{ij}\delta}{\frac{exp(c_{ij}/\tau_{ij})}{\sum_{k=1}^{N} exp(c_{ik}/\tau_{ik})}} \right] \\
\implies &\frac{\partial \tau_{ij}}{\partial c_{ij}} = \frac{\tau_{ij}}{c_{ij}} \left[1 - \frac{\tau_{ij}\delta \sum_{k=1}^{N} exp(c_{ik}/\tau_{ik})}{exp(c_{ij}/\tau_{ij})} \right] \\
\implies &\frac{\partial \tau_{ij}}{\partial c_{ij}} = \frac{\tau_{ij}}{c_{ij}} \left[1 - \tau_{ij}\delta \frac{exp(c_{ij}/\tau_{ij}) + \sum_{\substack{k=1 \\ k\neq j}}^{N} exp(c_{ik}/\tau_{ik})}{exp(c_{ij}/\tau_{ij})} \right] \\
\implies &\frac{\partial \tau_{ij}}{\partial c_{ij}} = \frac{\tau_{ij}}{c_{ij}} \left[1 - \tau_{ij}\delta \frac{exp(c_{ij}/\tau_{ij}) + \sum_{\substack{k=1 \\ k\neq j}}^{N} exp(c_{ik}/\tau_{ik})}{exp(c_{ij}/\tau_{ij})} \right] \\
\implies &\frac{\partial \tau_{ij}}{\partial c_{ij}} = \frac{\tau_{ij}}{c_{ij}} \left[1 - \tau_{ij}\delta \left(1 + \frac{\sum_{\substack{k=1 \\ k\neq j}}^{N} exp(c_{ik}/\tau_{ik})}{exp(c_{ij}/\tau_{ij})} \right) \right] \\
\implies &\frac{\partial \tau_{ij}}{\partial c_{ij}} = \frac{\tau_{ij}}{c_{ij}} \left[1 - \tau_{ij}\delta \left(1 + K \cdot exp(-\frac{c_{ij}}{\tau_{ij}}) \right) \right]
\end{aligned} \tag{6.11}$$

where $K = \sum_{\substack{k=1 \\ k\neq j}} exp(\frac{c_{ik}}{\tau_{ik}})$ is taken as a constant with respect to c_{ij}, that is, $\frac{\partial K}{\partial c_{ij}} = 0$.

If $N \to \infty$ or for very large N, we can safely assume

$$\frac{exp(c_{ij}/\tau_{ij}) + \sum_{\substack{k=1 \\ k\neq j}}^{N} exp(c_{ik}/\tau_{ik})}{exp(c_{ij}/\tau_{ij})} \simeq \frac{\sum_{\substack{k=1 \\ k\neq j}}^{N} exp(c_{ik}/\tau_{ik})}{exp(c_{ij}/\tau_{ij})} = K \cdot exp(-\frac{c_{ij}}{\tau_{ij}}) \tag{6.12}$$

Hence, Eqn. 6.11 reduces to,

$$\frac{\partial \tau_{ij}}{\partial c_{ij}} = \frac{\tau_{ij}}{c_{ij}} \left[1 - \tau_{ij}\delta \left(K \cdot exp(-\frac{c_{ij}}{\tau_{ij}}) \right) \right] \tag{6.13}$$

Solving the first-order nonlinear ordinary differential equation given by Eqn. 6.13, we get,

$$\tau_{ij} = \frac{c_{ij}}{\log(\delta \cdot K \cdot c_{ij} - k')} \tag{6.14}$$

where k' is the integral constant.

To find the value of k', we have to solve for the value of τ_{ij} at the endpoints of the cosine similarity line. It is to be remembered, τ_{ij} takes the value τ_{max} at $c_{ij} = -1$ and $c_{ij} = +1$ (Please refer to Sec. 6.3.3 in the main manuscript).

Solving, the above equation for the two above-mentioned cases, we get,

$$\begin{aligned} k'_{-} &= -\delta \cdot K - exp(-1/\tau_{max}) \\ k'_{+} &= -\delta \cdot K - exp(1/\tau_{max}) \end{aligned} \tag{6.15}$$

Varying the value of the constant in the range $[k'_{-}, k'_{+}]$, we get different curves with different slopes for different values of δ and K, as shown in the Fig. 6.2.

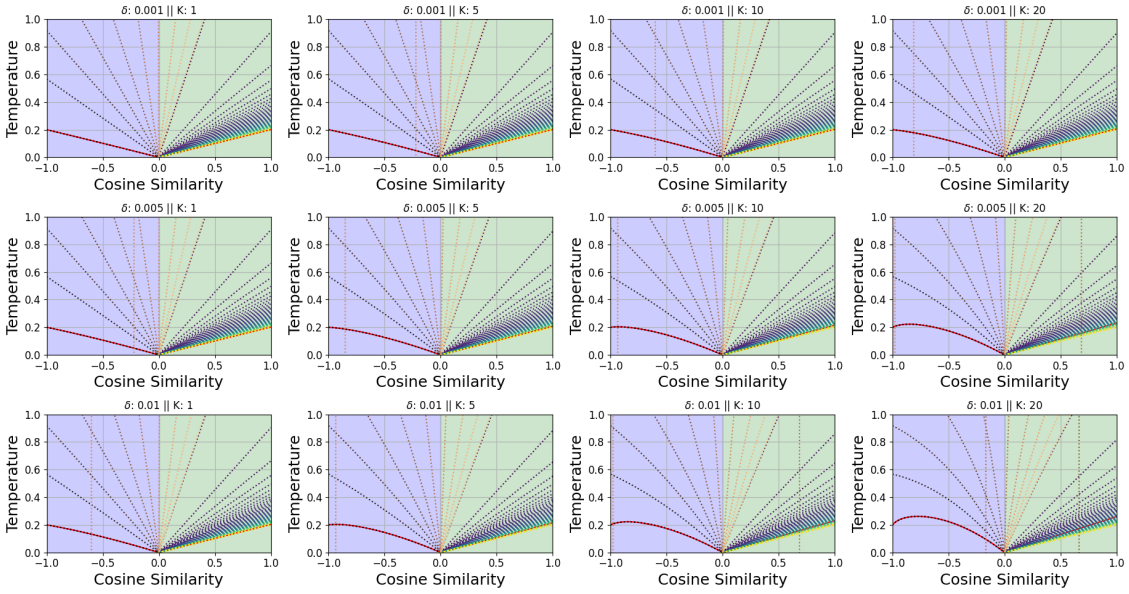


FIGURE 6.2: Plots of the solution of ODE in Eqn. 6.14 for different values of the integral constant, over different values of δ and K.

We can observe that the plotted curves in Fig. 6.2 show positive and negative gradients on the positive and negative half of the cosine similarity vs. temperature plane, respectively.

6.4.2 Proposed Temperature Scaling Function

Combining all the above philosophies we describe the framework proposed in this subsection. We use SimCLR (Chen et al., 2020a) as the baseline framework. For each pair in the InfoNCE loss function, the temperature hyper-parameter is replaced by a function given by Eqn. 6.16. To satisfy the characteristics of the slope of the temperature scaling function as derived in the section above, we adopt a cosine function of the cosine similarity as the temperature function, as shown in Eqn. 6.16.

$$\tau_{ij} = \tau_{min} + 0.5 \times (\tau_{max} - \tau_{min}) \times (1 + \cos(\pi(1 + c_{ij}))) \tag{6.16}$$

where τ_{max} and τ_{min} are the maximum and minimum limit of the temperature range, and c_{ij} denotes the cosine similarity of the pair (x_i, x_j).

Although we have used a smooth cosine function as the temperature scaling function, we present a further comparison of performance between different types of temperature functions such as linear and exponential functions in Sec. 6.5.4.1. We observe that we obtain similar performances using those functions. We can visualize from Fig. 6.3, that the cosine function does not violate the four criteria mentioned in Sec. 6.3.3.

For different τ_{max} and τ_{min} values, we obtain different temperature scaling functions as shown in Fig. 6.3. We also analyze how different temperature scaling schemes affect performance in Sec. 6.5.4.

Assigning lower temperatures to FN pairs pushes the constituent samples far apart. To reduce this effect, we can shift the minimum of the temperature function into the negative half of the cosine similarity vs. temperature plane. The result of this modification can be seen in Table 6.15. As the contribution corresponding to the FN pairs reduces in $\frac{\partial \mathcal{L}}{\partial z_i}$ in Eqn. 6.5, the performance also improves.

6.5 Experimental Details, Results and Analysis

We conducted extensive experiments on the ImageNet100, ImageNet1K, CIFAR10, and CIFAR100 datasets to empirically prove that the proposed temperature scaling framework outperforms the state-of-the-art SSL frameworks and also the recent temperature modulating framework MACL (Huang et al., 2023a). Similar to the previous chapter, we have used natural image datasets for pre-training. This allows us to transfer the representations learnt in the pre-training task to a variety of medical datasets. The unavailability of a large medical image dataset with diversity as natural image datasets like ImageNet limits the generalization capability of the model and subsequently limits the performance of the transferred tasks on different modalities. With appropriate fine-tuning, we can evaluate the quality of representations learnt in the pre-training step by using transfer learning tasks on both medical and natural image datasets. We compare the performance on transfer learning tasks with both SSL and supervised learning baselines.

6.5.1 Datasets

To study the effects of temperature in the self-supervised contrastive learning framework, we used the ImageNet1K (Deng et al., 2009) and ImageNet100 (Tian et al., 2020) datasets. We also used 3 small-scale datasets, namely, CIFAR10 (Krizhevsky, 2009) and CIFAR100 (Krizhevsky, 2009).Furthermore, we also study the effect of our proposed framework on the Long-tailed versions of the aforementioned datasets, which we term CIFAR10-LT and CIFAR100-LT. Like Chapter 5, in this chapter also we use medical data along with natural image data for our experiment to test the generalizability of the framework.

6.5.2 Implementation Details

In this section, we present the implementation details of both the pre-training and downstream transfer learning tasks of the proposed framework on both small-scale and large-scale datasets.

6.5.2.1 Pre-training Implementation Details

For the experiments on ImageNet1K and ImageNet100 datasets, we used a ResNet50 (He et al., 2016) backbone for all our experiments. The network parameters were optimized using a LARS optimizer with the square root learning rate scaling scheme as described in the SimCLR (Chen et al., 2020a) paper. For all our experiments we used a batch size of 256. The pre-training and the downstream tasks were run on a single 24GB NVIDIA A5000 GPU using the lightly-ai (Susmelj et al., 2020) library. To ensure faster training and prevent out-of-memory issues, automatic mixed precision (AMP) training was adopted. For the ImageNet100 dataset, we used a training duration of 200 epochs (approximately 40 hours), whereas for the ImageNet1k dataset, we pre-trained the encoder for 100 epochs (approximately 7 days).

The encoder used in the pre-training model for experiments on CIFAR10 and CIFAR100 is ResNet18 (He et al., 2016). The first convolutional layer in ResNet18 is replaced by a convolutional layer with a kernel of dimension 3×3 and the subsequent Max-pooling layer is removed. Similarly, for CIFAR10-LT and CIFAR100-LT, we used the aforementioned ResNet18, as well. However, for the experiments on Tiny-ImageNet, we use the original ResNet50 (He et al., 2016). The last fully connected layer from the ResNet network is removed for all experiments, and the output obtained from the ResNet encoder is fed into a 2-layer multi-layered perceptron (MLP) network called Projector. For the projector architecture, we follow the SimCLR (Chen et al., 2020a), where the Linear layers are followed by Batch Normalization (BN) (Ioffe and Szegedy, 2015) layers, with a ReLU (He et al., 2015) activation function in between the first BN layer and the second Linear layer. For the pre-training procedure, we use SGD optimizer (momentum= 0.9, weight decay factor= $5e-4$) with a learning rate of 0.06 and batch size of 128 for all datasets. For the balanced CIFAR and Tiny-ImageNet datasets, we run the optimization procedure for 200 epochs. For the long-tailed versions of the CIFAR datasets, we adopted 500 epochs of training (Kukleva et al., 2023).

For the evaluation stage, we adopt a kNN classifier. For the balanced CIFAR and Tiny-ImageNet datasets, we used a kNN classifier with $k = 200$, with weights based on cosine similarity. For long-tailed versions of the CIFAR datasets, we used a k value of 1 and 10 with weights based on $l2$-distance. All the training and inference were run on a 16GB NVIDIA P100 GPU. Since the proposed framework is based on InfoNCE, the computation overhead is the same as contemporary frameworks such as SimCLR (Chen et al., 2020a), MoCov2 (Chen et al., 2020c), etc.

Augmentations During the pre-training stage, two augmented versions of each input image are generated and used as positive pairs. For the augmentations, we follow the augmentation strategy of SimCLR (Chen et al., 2020a).

6.5.2.2 Transfer Learning Implementation Details

We fine-tuned the models pre-trained on the ImageNet1K (Deng et al., 2009) dataset for 50 epochs using an SGD optimizer. We used class weights to mitigate the effect of imbalance in all datasets, except the natural image datasets. For the MURA, ISIC2016, and MHIST datasets, we used positive class weights of 0.7097, 4.24, and 2.45, respectively. For the Chaoyang dataset, the class weights used were 1.264, 1.667, 1.0, and 2.114 for the 4 classes. For all the experiments, a batch size of 128 was used. For the multiclass and binary classification tasks, we used a learning rate of 0.1 and 1.0, respectively, and a multistep decay scheduler with a decay by a factor of 0.1 at the 30th and 40th epochs.

6.5.3 Comparative Results and Analysis

In this section, we present the comparative results of the proposed framework on small-scale datasets (CIFAR10, CIFAR100) in Sec. 6.5.3.1, and large-scale datasets (ImageNet100, ImageNet1k) in Sec. 6.5.3.3. We also present comparative results on transfer learning tasks on both medical and natural image datasets to evaluate the quality of learned representations.

6.5.3.1 Results on Small Scale Datasets

In this section, we present the performance of our proposed framework on CIFAR10 and CIFAR100 (Table 6.1) datasets. We have reported the best results obtained with $\tau_{min} = 0.07$ and $\tau_{max} = 0.2$ for comparison in Table 6.1. From our ablations on CIFAR datasets (Sec. 6.5.4), we observed that the vanilla DySTreSS works better than the shifted versions. On the long-tailed datasets (Table 6.2 and 6.3), the DySTreSS outperforms Kukleva et al. (2023) after 2000 epochs of pre-training.

TABLE 6.1: Comparison with SOTA SSL frameworks on CIFAR10 and CIFAR100 datasets.

Framework	Temp. Scaled	CIFAR10	CIFAR100
SimCLR (Chen et al., 2020a)	×	83.65	52.32
MoCoV2 (Chen et al., 2020c)	×	83.9	54.01
SimCLR+DCL (Yeh et al., 2022)	×	84.4	56.02
MACL (repro.) (Huang et al., 2023a)	✓	84.85	56.15
DySTreSS (Proposed)	✓	**85.68**	**56.57**

TABLE 6.2: Comparis on long-tailed CIFAR datasets.

Framework	CIFAR10-LT	CIFAR100-LT
Kukleva et al. (2023)	62.91	30.20
DySTreSS (Proposed)	**64.98**	**31.71**

TABLE 6.3: Comparison on ImageNet100-LT datasets.

Framework	ImageNet100-LT
Kukleva et al. (2023)	45.3 (repro.)
DySTreSS (Proposed)	**46.1**

On the CIFAR-10 dataset, the proposed framework outperforms the recent state-of-the-art (SOTA) framework MACL (Huang et al., 2023a) along with other contrastive SSL frameworks. On the CIFAR-100 dataset, DySTreSS outperforms SimCLR and MACL by 1.77% and 1.39%, respectively.

We also present the comparison of the performance of our proposed framework with two non-contrastive frameworks, DINO (Caron et al., 2021) and WMSE (Ermolov et al., 2021) on CIFAR datasets in Table 6.4. All experiments were done with a batch size of 256 and trained for 200 epochs.

TABLE 6.4: Comparison with DINO and WMSE on CIFAR datasets

Methods	CIFAR10	CIFAR100
DINO (Caron et al., 2021)	84.02	46.79
WMSE (Ermolov et al., 2021)	85.52	52.72
DySTreSS (Proposed)	**85.68**	**56.57**

6.5.3.2 Results on Long-Tailed Datasets

On the long-tailed versions of the CIFAR datasets, the proposed framework improves upon the baseline SimCLR by more than 1%, as seen in tables 6.5 and 6.6. ✓ and ✗ in the "Temp. Scaled" column indicates if the corresponding framework uses temperature scaling. We also see in Table 6.2 and 6.3, that our proposed method performs better than Kukleva et al. (2023) when pre-trained for 2000 epochs, as per the pre-training configuration in Kukleva et al. (2023). For the experiments on ImageNet100-LT, we used a batch size of 256 for both Kukleva et al. (2023) and DySTreSS.

TABLE 6.5: Comparison with state-of-the-art SSL frameworks on CIFAR10-**LT** dataset. DySTreSS and DySTreSS* both have $\tau_{max} = 0.2$, but the values of τ_{min} are 0.07 and 0.1, respectively.

Framework	Temp. Scaled	Accuracy	
		1-NN	10-NN
SimCLR (Chen et al., 2020a)	×	57.12	55.29
DySTreSS (Proposed)	✓	**58.36**	56.40
DySTreSS* (Variation of Proposed)	✓	58.34	**56.54**

TABLE 6.6: Comparison with state-of-the-art SSL frameworks on CIFAR100-**LT** dataset. DySTreSS and DySTreSS* both have $\tau_{max} = 0.2$, but the values of τ_{min} are 0.07 and 0.1, respectively.

Framework	Temp. Scaled	Accuracy	
		1-NN	10-NN
SimCLR (Chen et al., 2020a)	×	28.27	26.18
DySTreSS (Proposed)	✓	**29.43**	27.10
DySTreSS* (Variation of Proposed)	✓	28.82	**27.32**

6.5.3.3 Results on Large Scale Datasets

In this section, we present the comparative results of the proposed DySTreSS framework on large-scale datasets ImageNet100 and ImageNet1k. We compare with both contrastive and non-contrastive frameworks. We observe that the proposed framework outperforms the contemporary state-of-the-art SSL frameworks.

Comparison with Contrastive Frameworks

We consider the vanilla SimCLR as the baseline for our proposed temperature scaling framework, hence all the experiments on ImageNet100 and ImageNet1k were conducted on the SimCLR framework. We also compare our work with the state-of-the-art contrastive learning frameworks. In Tab. 6.7, results of the vanilla DySTreSS framework and a shifted version of the same (Sec. 6.5.4.3) with shift and scale parameters $\triangle s$ and k, respectively are also presented.

As the implementation is done using the lightly-ai (Susmelj et al., 2020) library, we use the benchmark results provided by the library on the ImageNet1K dataset for the comparison. The results shown in Table 6.8 are for 100 epochs of pre-training on ImageNet1K with a batch size of 256.

The values of τ_{min} and τ_{max} were set to 0.1 and 0.2, respectively. ✓ and ✗ in the "Temp. Scaled" column indicates if the corresponding framework uses temperature scaling. SimCLR was used as the baseline in DCL, MACL, and DySTreSS.

TABLE 6.7: Comparison with state-of-the-art Contrastive SSL frameworks on the ImageNet100 dataset.

Framework	Temp. Scaled	Lin. Eval. Acc. Top-1	Top-5
SimCLR (Chen et al., 2020a)	×	75.54	93.06
MoCoV2 (Chen et al., 2020c)	×	76.80	94.34
DCL (Yeh et al., 2022)	×	77.38	94.01
MACL (Huang et al., 2023a)	✓	78.28	94.25
DySTreSS (Proposed)	✓	**78.78**	**94.70**

TABLE 6.8: Comparison with state-of-the-art Contrastive SSL frameworks on the ImageNet1K dataset

Framework	Temp. Scaled	Lin. Eval. Acc. Top-1	Top-5
SimCLR (Chen et al., 2020a)	×	63.2	85.2
DCL (Yeh et al., 2022)	×	65.1	86.2
DCLW (Yeh et al., 2022)	×	64.2	86.0
MACL (Huang et al., 2023a)	✓	64.3	-
DySTreSS (Proposed)	✓	**65.21**	**86.55**

We observe that the proposed framework outperforms the contemporary state-of-the-art SSL methods on linear probing evaluation. The proposed framework also outperforms the recent state-of-the-art framework MACL (Huang et al., 2023a) which also adopts a temperature-modifying approach on top of the contrastive learning SimCLR framework. Furthermore, we also applied the proposed DySTreSS framework on the SimCLR+DCL framework with the base configuration $\tau_{min} = 0.1$ and $\tau_{max} = 0.2$, and achieved an improvement of 0.14% over the Top-1 accuracy reported in Table 6.7.

Comparison with Non-Contrastive Frameworks

In this section, we present the comparison of the proposed DySTreSS framework with the non-contrastive SSL frameworks on both ImageNet100 and ImageNet1k datasets. For the comparison on the ImageNet1k dataset in Table 6.9, we ran the DySTreSS framework for 100 epochs with a batch size of 256. The comparative results were available from the lightly-ai (Susmelj et al., 2020) benchmark results.

For the ImageNet100 dataset, we ran pre-training for 400 epochs with a batch size of 128 as we used the same training conditions as in Zhang et al. (2022a) and Zhang et al. (2022b), for a fair comparison. From the results presented in Table 6.10 and obtained on the ImageNet100 dataset, we can see that our proposed method outperforms the state-of-the-art methods like DINO (Caron et al., 2021), WMSE (Ermolov et al., 2021), Zero-CL (Zhang et al., 2022a), and ARB (Zhang et al., 2022b). Linear evaluation task accuracy values for WMSE on the ImageNet100 dataset for 400 epochs pre-training are taken from Zhang et al. (2022a) and Zhang et al. (2022b).

TABLE 6.9: Comparison with state-of-the-art Non-contrastive SSL frameworks on ImageNet1K dataset (Here, B. Twins stands for Barlow Twins)

Framework	Temp. Scaled	Lin. Eval. Acc.	
		Top-1	Top-5
BYOL (Grill et al., 2020)	N/A	62.4	82.7
B. Twins (Zbontar et al., 2021)	N/A	62.9	84.3
VicReg (Bardes et al., 2022a)	N/A	63.0	85.4
DySTreSS (Proposed)	✓	**65.21**	**86.55**

TABLE 6.10: Comparison of the proposed method with Non-contrastive SSL frameworks DINO, WMSE, Zero-CL, and ARB on the ImageNet100 dataset on Linear Evaluation task.

Frameworks	Proj. Dim #	Linear Eval. Acc.	
		Top - 1	Top - 5
Barlow Twins (Zbontar et al., 2021)	2048	78.62	94.72
VICReg (Bardes et al., 2022a)	2048	79.22	95.06
ZeroICL (Zhang et al., 2022a)	256	78.02	95.61
ZeroFCL (Zhang et al., 2022a)	2048	79.32	94.94
ZeroCL (Zhang et al., 2022a)	2048	79.26	94.98
WMSE (Ermolov et al., 2021)	256	69.06	91.22
ARB (Zhang et al., 2022b)	2048	79.48	95.51
DINO (Caron et al., 2021)	256	74.84	92.92
BYOL (Grill et al., 2020)	4096	80.09	94.99
LogDet (Zhang et al., 2024a)	2048	80.38	95.45
DySTreSS (Proposed)	2048	**81.24**	**95.64**

6.5.3.4 Comparison of Performance in Transfer Learning Setting

In this section, we present the comparison of transfer learning performance on both medical and natural image datasets with different degrees of imbalance. These tasks include both binary and multi-class classification tasks. The encoders were initialized with weights obtained after pre-training for 100 epochs on the ImageNet1K dataset. We also provide the supervised baseline performance as a reference. Implementation details are discussed in detail in Section 6.5.2.2.

Transfer Learning Performance on Medical Image Datasets
We choose four medical image datasets, MURA, Chaoyang, ISIC2016 Lesion Classification, and MHIST datasets. The MURA, ISIC2016 Lesion classification and MHIST consist of the binary classification task. The Chaoyang consists of multi-class classification tasks.

From the results presented in Table 6.11, we can see that the proposed method outperforms SimCLR (Chen et al., 2020a) on all 4 medical datasets. The proposed method also outperforms the current state-of-the-art self-supervised contrastive learning algorithm like

DCL (Yeh et al., 2022) on 2 out of 4 datasets. It can also be seen that the performance of our proposed framework is close to the supervised baseline.

TABLE 6.11: Performance comparison of the proposed method (DySTreSS) with contemporary self-supervised contrastive state-of-the-art methods on transfer learning tasks on medical image datasets. The results of the supervised learning baseline are also provided here for reference.

Datasets	SimCLR	DCL	DySTreSS	Supervised
MURA	81.81	81.70	**82.27**	82.10 (Nauta et al., 2023b,a)
Chaoyang	83.22	83.12	**83.26**	83.50 (Galdran et al., 2023)
ISIC2016	85.49	**86.02**	85.75	85.50 (ISDIS, 2016)
MHIST	83.62	**85.26**	83.98	86.90 (Springenberg et al., 2023)

Transfer Learning Performance on Natural Image Datasets
We also choose three natural image datasets, CIFAR10, CIFAR100 and Flowers. The Flowers, CIFAR10, and CIFAR100 consist of the multi-class classification tasks.

From the results presented in Table 6.12, we can see that the proposed method outperforms SimCLR (Chen et al., 2020a) on all the 3 natural datasets. The proposed method also outperforms the current state-of-the-art self-supervised contrastive learning algorithm like DCL (Yeh et al., 2022) on 3 out of 3 datasets. It can also be seen that the performance of our proposed framework is close to the supervised baseline.

TABLE 6.12: Performance comparison of the proposed method (DySTreSS) with contemporary self-supervised contrastive state-of-the-art methods on transfer learning tasks on natural image datasets. The results of the supervised learning baseline are also provided here for reference.

Datasets	SimCLR	DCL	DySTreSS	Supervised
CIFAR10	96.93	97.09	**97.14**	97.50 (Grill et al., 2020)
CIFAR100	82.99	83.03	**83.96**	86.40 (Grill et al., 2020)
Flowers	93.82	94.11	**96.76**	97.60 (Grill et al., 2020)

6.5.4 Ablation Studies

6.5.4.1 Effect of Different Temperature Functions

In this section, we present the performance of different temperature functions, namely linear and exponential on CIFAR datasets. From Table 6.13, we see that all the temperature functions perform similarly to the cosine function, if not better. All the temperature functions satisfy the conditions of positive and negative slopes on the positive and negative half of the temperature vs. cosine similarity plane, respectively. All experiments were run for 200 epochs, and the results reported are for 200-NN Top-1 accuracy. We observe that as long as the temperature scaling function follows the basic principle of having positive and negative slopes in the positive and negative half of the temperature vs. cosine similarity plane, application of the temperature scaling function improves the performance of the baseline SimCLR framework on the downstream task.

TABLE 6.13: Comparison of performance on CIFAR datasets for different temperature functions

Function	CIFAR10	CIFAR100
Cosine	85.85	56.57
Linear	85.74	56.78
Exponential	85.81	56.47

6.5.4.2 Effect of Temperature Range

To find the optimal range of temperature, we conduct pre-training with several temperature ranges on the ImageNet100 dataset as given in Table 6.14 and find that the values $\tau_{min} = 0.1$ and $\tau_{max} = 0.2$ gives the best performance. The temperature functions for different temperature ranges are also shown in Figure 6.3. We observe that setting the value of τ_{max} towards high degrades performance, whereas setting both τ_{max} and τ_{min} towards low also degrades performance. The reason for such behaviour can be understood from the uniformity and tolerance plots in Fig. 6.4.

When both the τ_{max} and τ_{min} are low, the repulsion between samples in false negative pairs is amplified, resulting in increased uniformity. A low temperature also results in a high tolerance due to increased attraction forces between samples in a positive pair. Keeping τ_{min} fixed, if we increase the τ_{max} the uniformity decreases due to a decrease in the repulsion between samples in the false negative pairs and then increases again due to decreased attraction between the samples in a positive pair.

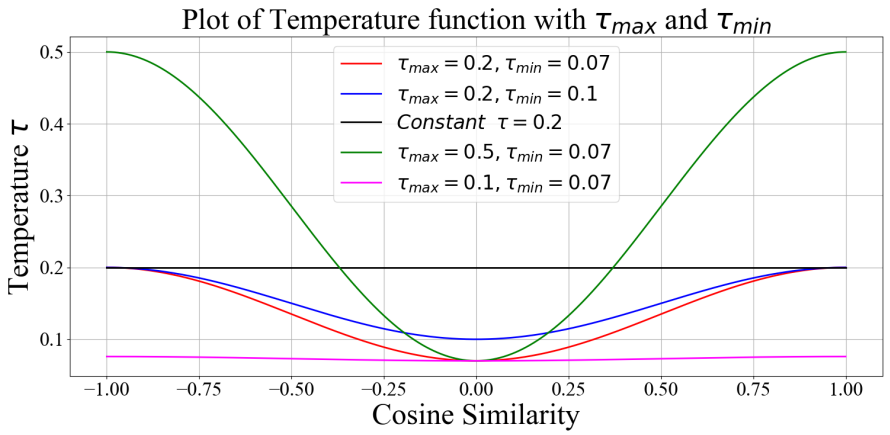


FIGURE 6.3: Temperature functions for different τ_{max} and τ_{min}.

TABLE 6.14: Ablation of DySTreSS on different temperature ranges on ImageNet100 dataset.

τ_{min}	τ_{max}	20NN Acc.		Lin. Eval. Acc.	
		Top-1	Top-5	Top-1	Top-5
0.07	0.1	67.18	87.82	77.28	94.16
0.07	0.5	67.88	88.22	76.34	93.72
0.07	0.2	71.46	89.42	78.46	94.4
0.1	0.2	72.00	89.76	78.76	94.7

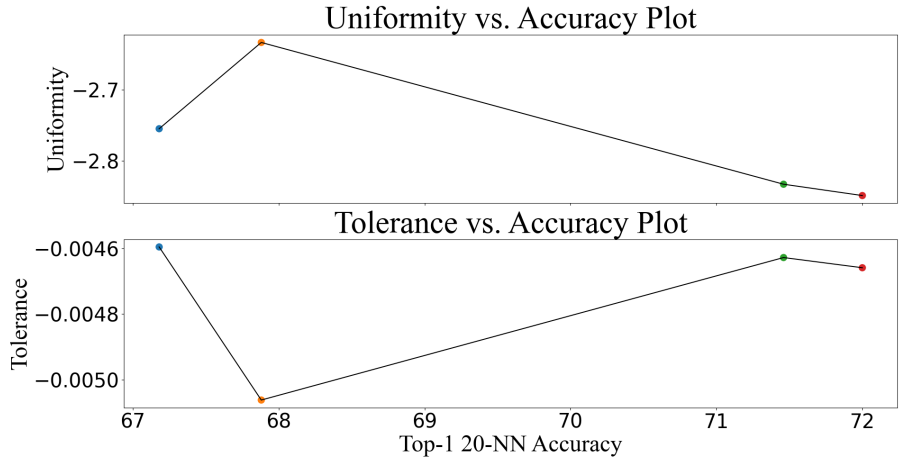


FIGURE 6.4: Plot of Uniformity and Tolerance vs. 20NN Top-1 acc. shown in Table 6.14.

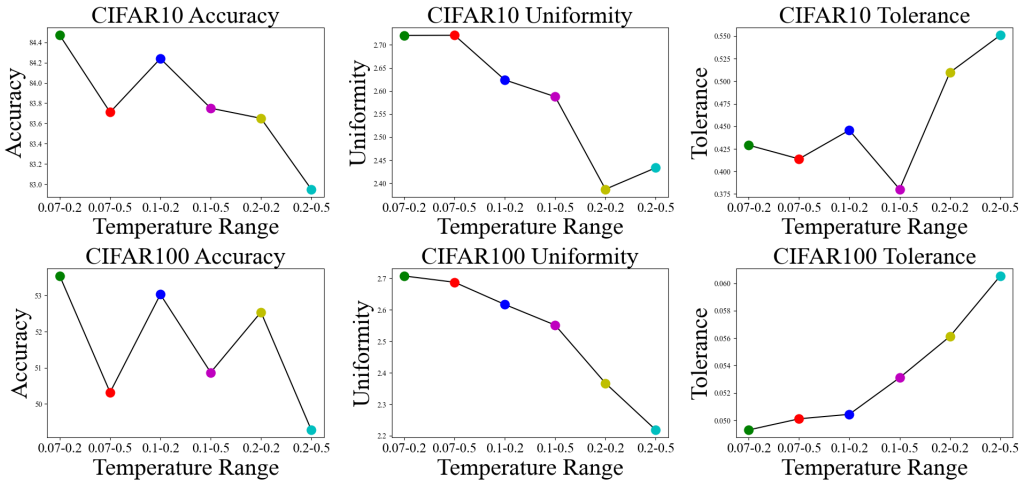


FIGURE 6.5: Plot of Accuracy on CIFAR10 (top) and CIFAR100 (bottom) datasets for different temperature ranges.

We also observe similar behaviour of uniformity and tolerance on the CIFAR10 and CIFAR100 datasets. It is evident from Fig. 6.5, that an increase in temperature causes a decrease in the magnitude of the uniformity metric. Whereas, if we lower τ_{min} but keep the same τ_{max}, the general trend shows that the magnitude of the uniformity increases and tolerance decreases. This conforms with our theory in Sec. 6.3.2. On the contrary, for CIFAR100, due to the presence of *more true negative pairs* than CIFAR10, increasing temperature inhibits the uniformity (tolerance) from increasing (decreasing) sufficiently to improve performance. Hence, for the same τ_{min}, the performance was better for a lower τ_{max}.

6.5.4.3 Effect of Shifted Temperature Profiles

Similar to the previous ablation study, we look to find the optimal temperature profile over the cosine similarity spectrum by shifting the minima of the temperature profile to the left or right as depicted in Figure 6.6. When applying a shift ($\triangle c$) to the minima at $c = 0.0$, we also scale the temperature profile by k to ensure that the temperatures at the extremities of the cosine similarity spectrum ($c = \pm 1.0$) remain at τ_{max} according to our assumptions in Sec. 6.3.2. The algorithm for computing the values of the shifted temperature functions is presented in Alg. 6.1.

Algorithm 6.1: Shifted Temperature Functions

Data: $\tau_{max}, \tau_{min}, \Delta\tau = \tau_{max} - \tau_{min}, \Delta c, k_s$
Input: $c_{ij} \rightarrow$ Cosine Similarity of (x_i, x_j)
if $(\Delta c \leq 0 \wedge c_{ij} \leq -\Delta c) \vee (\Delta c \geq 0 \wedge c_{ij} \geq -\Delta c)$ **then**
  $\tau_{ij} = \tau_{min} + \frac{\Delta\tau}{2}(1 + \cos(\frac{\pi}{k_s}(\Delta c + c_{ij})))$
else
  $\tau_{ij} = \tau_{max}$

We primarily use three different shifted versions of the temperature profile for the ablations on the ImageNet100 dataset. In the first version, Shifted Minima Ver. 1, we shift the

minima towards the right half-plane. In the second version, Shifted Minima Ver. 2, we shift the minima towards the left half plane. In the last and third versions, we shift the temperature profile entirely in the left half plane and keep the temperature constant in the right half plane. 'SMvx' denotes 'Shifted Minima Version x'. 'Ver. xa' and 'Ver. xb' denotes two shifts of -0.2 and -0.4 from the origin. In addition to the shift, we also apply appropriate scaling such that the extremities have maximum temperature. The algorithm for calculating the temperature for the shifted temperature profile is given in Alg. 6.1. We observe from the results presented in Table 6.15, that a shift of $\triangle c = -0.4$ from $c = 0.0$ and a scaling of $k_s = 0.7$ yields the best linear evaluation performance on the ImageNet100 dataset.

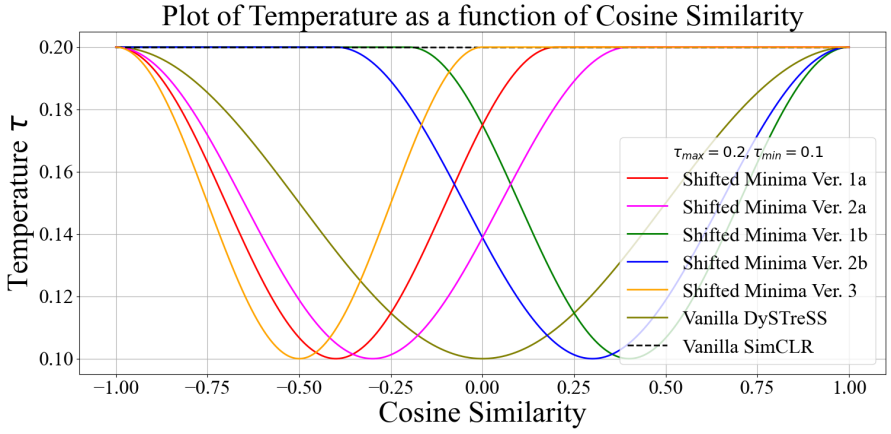


FIGURE 6.6: Plot of shifted versions of Temperature functions.

TABLE 6.15: Ablation on different temperature profiles on ImageNet100 dataset.

shift	scale	20-NN Accuracy		Lin. Eval. Acc.	
		Top-1	Top-5	Top-1	Top-5
0.0	0.5	71.66	90.16	78.46	94.86
0.2	0.6	71.58	89.14	78.56	94.38
0.4	0.7	71.58	89.58	78.46	94.54
-0.2	0.6	72.38	89.99	78.78	94.59
-0.4	0.7	71.58	90.16	78.82	94.76

In Fig. 6.7, we present the accuracy, uniformity, and tolerance values for different temperature functions given in Fig. 6.6 for the CIFAR10 and CIFAR100 datasets. We observe that shifting the minimum of the temperature function influences different samples and consequently changes the structure of the feature space and performance accordingly. For the CIFAR10 dataset, we can observe that a shift of -0.2 is better than a shift of -0.4, while the reverse is true for the CIFAR100 dataset. However, none of the configurations yields better results than the vanilla version with no shift. A lower temperature towards $c_{ij} = -1$ increases uniformity, as evident from the difference in uniformity

between SMv1b and SMv2b or SMV1a and SMv2a, while the reverse is true for tolerance. The drop in performance is primarily because a constant temperature in the range $[-\tau_{shift}, 1.0] \mid (\tau_{shift} \in \{-0.2, -0.4\})$ caused by the shift results in decreased repulsion of the hard true negative samples, delaying convergence.

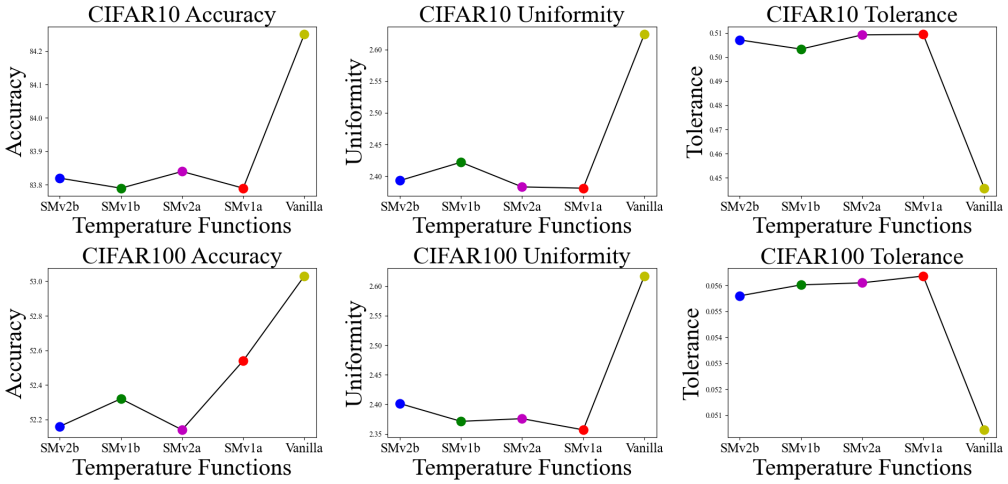


FIGURE 6.7: Plot of Accuracy, Uniformity, and Tolerance with a shift in minima for CIFAR10 (top) and CIFAR100 (bottom) dataset. The colour codes are matched to the curves in Fig. 6.6.

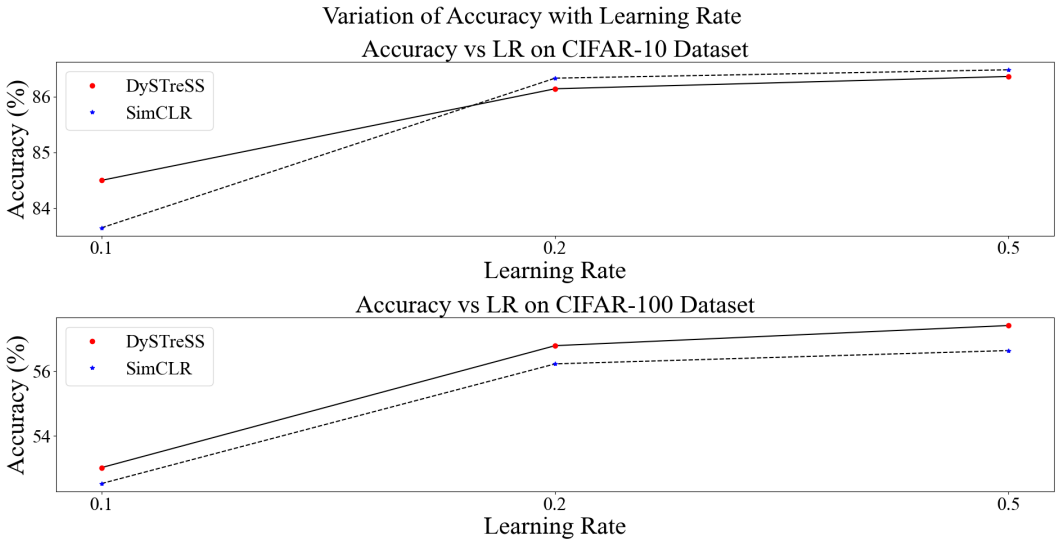


FIGURE 6.8: Plot of Accuracy with change in Learning Rate for the datasets CIFAR10 (top) and CIFAR100 (bottom).

6.5.4.4 Effect of Learning Rate

As the temperature is decreased, the displacement gradients (Eqn. 6.5) increase, increasing the magnitude of fluctuations from false negative pairs. A low learning rate plays a crucial role in this scenario in smoothening out the fluctuations. On the contrary, at a high

learning rate, the fluctuations are amplified and should degrade the performance. However, from Fig. 6.8, we observe this effect for CIFAR10 only, as the number of false negative pairs in a batch is greater than that in CIFAR100.

6.6 Conclusion

In this work, we identified a specific category of pairs in self-supervised contrastive learning, and analyzed the effect of temperature on such pairs in optimizing InfoNCE loss. We observed that by varying the temperature as a function of the cosine similarity values of the feature vectors of all pairs, we can control the dynamics of the optimization process and improve the performance of the baseline method, SimCLR. Through extensive experiments, we show that the proposed framework improves performance over the baseline and state-of-the-art algorithms. Finally, this work lays the foundation for further research into the working principle and dynamics of the InfoNCE loss function.

Generally, the pretext and downstream tasks differ in their objectives. Thus, the representations learnt in the pretext task, although being adapted to the data distribution, may not be optimal for the downstream task. Hence, we propose a pretext task which learns representations that are aligned to the downstream task, and in the next chapter we look to minimize the disparity between the pretext and downstream tasks.

Chapter 7

Self-Supervised Learning for Medical Image Segmentation using Prototype Aggregation

7.1 Introduction

Semantic Segmentation is one of the critical applications in computer vision. Applications of semantic segmentation to medical image analysis for assisting medical personnel in disease diagnosis are also plenty. For efficient and reliable analysis of medical images, contemporary deep-learning methods require large-scale datasets annotated by expert medical personnel. However, obtaining annotated high-quality medical image datasets is time-consuming and labour-intensive, unlike natural image datasets. The few-shot learning paradigm has gained popularity among researchers to avoid the scarcity of large-scale datasets in medical image analysis.

One of the first pioneering works in Few Shot Segmentation (FSS) was presented in Shaban et al. (2017), where a conditioning branch was used to predict weights, which serves as classifier weights for the query image feature obtained from the segmentation branch. The idea was extended in Rakelly et al. (2018a) using sparse positive and negative support pixels. In Rakelly et al. (2018b), a guided network is used to utilize information from the latent representation from the FSS task to segment query pixels. This work was further improved and extended in Zhang et al. (2020); Siam and Oreshkin (2019); Siam et al. (2019) and Bhunia et al. (2019). Dong and Xing (2018) presents one of the first works in the Prototypical Learning paradigm, by fusing prototypes from support images with the query image features using similarity scores. A leap in the paradigm of prototype learning was shown in Wang et al. (2019), where the prototype alignment strategy was introduced for maintaining cyclic consistency between the ground truth and the predicted segmentation mask inducing a regularizing effect in training. Instead of altering the input structure as in Dong and Xing (2018) and Rakelly et al. (2018a), the authors in Zhang et al. (2020) adopted a separate segmentation guidance framework based on similarity. Liu et al. (2020) argue that the previous prototype-based methods do not take into account

the various appearance of different parts in an object and propose a prototype-based part-aware framework to capture rich and fine-grained features. Yang et al. (2020) also pointed out that the primary disadvantage of existing prototype-based methods is the pooling operations which destroy the spatial layout information of the objects, and thereby proposed a prototype mixture model to solve the semantic ambiguity in prototype-based models.

To preserve the spatial correspondence between support and query image pixels, Liu et al. (2022b) uses a partial optimal transport-based matching. A multi-level variation of the same was done in Wang et al. (2020). Li et al. (2021a) and Fan et al. (2022) also aim to solve the same problem. Zhang et al. (2021a) and Liu et al. (2022a) aim to capture the intrinsic details to improve segmentation quality. He et al. (2021) and Chen et al. (2023a) attempts to reduce testing bias in the FSS setting. An attempt to improve the discrimination between similar classes is presented in Okazawa (2022).

In the approaches discussed in the previous chapters, we observed that the pretext and downstream tasks are not the same. For example, in Chapter 3, the pretext task is jigsaw puzzle solving or a multi-class classification task, but the downstream task is knee injury classification or a binary classification task. Again, in Chapter 4, 5 and 6, the pretext task is a paired embedding-based instance discrimination contrastive learning task, whereas the downstream task is a multi-class classification task. While self-supervised pre-training aids in the appropriate parameters to be learnt and are better aligned to the target dataset than weights pre-trained on other datasets, the semantic representations learnt depend on the pretext task. It can also not be denied that the model trained using a pretext task will learn representations which are not relevant to the downstream task. Hence, in this chapter, we adopt the one-shot learning approach for learning to segment organs from MR query scans, from a limited number of given support MR slices and their corresponding ground truth segmentation masks.

The application of self-supervised learning frameworks in segmentation, although limited, also follows two paths as discussed above. One where the pre-training task is different from the downstream task of segmentation, and the other where both are the same. Works like Guizilini and Ramos (2013); Singh et al. (2018); Ji et al. (2019); Chen et al. (2019); Zhu et al. (2020); Ouali et al. (2020); Hoyer et al. (2021); Gao et al. (2022b) use a pre-training stage to learn representations from the base dataset and then utilize the representation for a downstream semantic segmentation task. Gansbeke et al. (2021) uses unsupervised saliency to generate object proposals and then optimizes a contrastive learning objective on the features obtained from the proposals to learn representations for semantic segmentation. For the second type, pseudo-masks are used as the segmentation masks in the pre-training stage. The above strategy is adopted in Ouyang et al. (2020); Araslanov and Roth (2021); Ouyang et al. (2022) and Amac et al. (2022). In Araslanov and Roth (2021), the authors use the output masks of a momentum update net as target pseudo-masks. In Ouyang et al. (2020, 2022), the pseudo-masks are generated using the Felzenszwalb algorithm (Felzenszwalb and Huttenlocher, 2004) and the model using a few-shot learning strategy, the query image is an augmented version of the support image itself.

From the above discussion, we see that depending on the pipeline, few-shot segmentation frameworks can be primarily of two types: prototype feature learning and affinity learning

(Li et al., 2021a). Prototype feature learning consists of constructing prototypes utilizing the support image and the support mask information. Each prototype represents a defined spatial region in the support image. These prototypes are used to find pixels in the query image which are similar to them and are scored accordingly to segment it into foreground and background. Prototype-based features are more robust to noise than pixel-based features (Li et al., 2021a). Prototypical methods also drop spatial information, which is important when the variation between support and query images is considerably significant (Li et al., 2021a). Prototypical methods also are responsible for losing discriminability because of the masked pooling process to generate prototypes (Li et al., 2021a). To address this issue, we generate prototypes for both the foreground and background pixels, which preserve the contextual spatial information required for effective discrimination between the foreground and background pixels. To prevent loss of information, we adopt a correlation-weighted prototype aggregation approach such that the information of all the prototypes corresponding to foreground or background is present in the aggregated prototype.

In PANet (Wang et al., 2019), a prototypical alignment-based strategy was proposed, wherein the masked support image embedding is mapped to the feature space, and the query mask is predicted by matching the query prototypes to the nearest prototype in the embedding space. However, PANet resorts to a global masked pooling operation, which is not suitable for medical image segmentation, as it can result in the loss of spatial orientation information. In ALPNet (Ouyang et al., 2022), which is also a prototype-based framework, a local prototype-based approach is adopted to preserve local information using an adaptive local prototype pooling framework. However, such an approach ignores global contextual information.

In this chapter, we propose a prototype-based one-shot learning framework for the segmentation of organs in MR scans. We take a correlation-based aggregation approach to generate dynamic prototypes to encode spatial context and assign probabilities to the prototypes based on correlation-based matching. In addition to the aforementioned framework design, we also utilize prior domain information to further reduce the effect of false positives in the final downstream task by using *quadrant masking* scheme. We provide extensive experimental evidence on two datasets on abdominal magnetic resonance imaging and computed tomography showing the efficacy of the proposed simple but potent method.

The rest of the paper is organized as follows: Sec. 7.2 describes the preliminaries related to this chapter. Sec. 7.3 deals with the motivation behind the proposed self-supervised one-shot segmentation framework. Next, in Sec. 7.4 we discuss in detail the working principles of our proposed framework. After that, we provide the implementation details and the results obtained from our extensive experimentation including ablation studies in Sec. 7.5. Finally, we end the technical section with a brief conclusion in Sec. 7.6.

7.2 Preliminaries

Few Shot Learning

In the few-shot learning framework, the dataset is split into two parts, training dataset $\mathcal{D}_{train}$ and testing dataset $\mathcal{D}_{test}$. In both training and testing datasets, each sample consists of the input and the associated ground truth, (I, M). In our work, I and M correspond to slices from the abdominal MR scans and the associated superpixel pseudo-masks generated in the pre-training stage, respectively. Whereas, in the evaluation or testing stage, the original ground truth masks for each organ are used with the MR slices. Furthermore, no overlap should be present between the classes present in $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$.

In the few-shot learning framework, we need to consider two sets of data, the *Support* set $\mathcal{S}$ and the *Query* set $\mathcal{Q}$. The *Support* set $\mathcal{S}$ consists of the tuple $\{I_s^i, M_s^i(l)\}_{i=1}^k$, where I_s^i is the i-th sample in the Support set with the segmentation mask $M_s^i(l)$ for the class l, where l belongs to the set of novel classes available during the testing phase. The primary objective is to learn an approximate function f which takes as input the support set $\mathcal{S}$ and the query image I_q and predicts the binary mask $\hat{M}_q$ of the unseen classes in I_q, denoted by the support mask $M_s(l)$.

The support set $\mathcal{S}$ is a subset of $\mathcal{D}_{train}$. During training, the input to the model is $(\mathcal{S}, I_q)$. Such a pair is called an *episode*. If during training, the value of k is 1, that is, we use only a single image in the support set, then the learning is known as one-shot learning, which we adopt in this work. If $k > 1$, it is known as few-shot learning. If the number of classes is $\mathcal{C}$, then we call it $\mathcal{C}$-way k-shot learning.

7.3 Motivation

The primary motivation for the framework proposed in this chapter stems from the observation that the pretext task and the downstream task often differ in their objectives. We have already discussed previously that the pretext task controls the quality and type of representations that will be learnt and transferred to the downstream task. Hence, it plays a significant role in determining the adaptability of the transferred features. Thus, intuitively, the representations learnt in the pretext task should be aligned with the nature of representations required in the downstream task. Consequently, this can also help reduce fine-tuning efforts and computational overhead. However, only a few studies have been done to minimize the pretext and downstream task disparity. Among the recent studies, ALPNet (Ouyang et al., 2022) and CRAPNet (Ding et al., 2023) are the noteworthy few-shot self-supervised segmentation frameworks which are also driven by the same motivation. In this chapter, we propose a novel correlation-weighted prototype aggregation-based self-supervised one-shot segmentation framework to learn representations which can be utilised to segment abdominal organs from MR scans without downstream fine-tuning.

7.4 Proposed Framework

In this section, we discuss the proposed framework. This section primarily consists of two parts: (1) Pretext task framework which describes the steps followed for pre-training the one-shot segmentation model, and (2) Downstream task framework which discusses the steps for one-shot segmentation without fine-tuning.

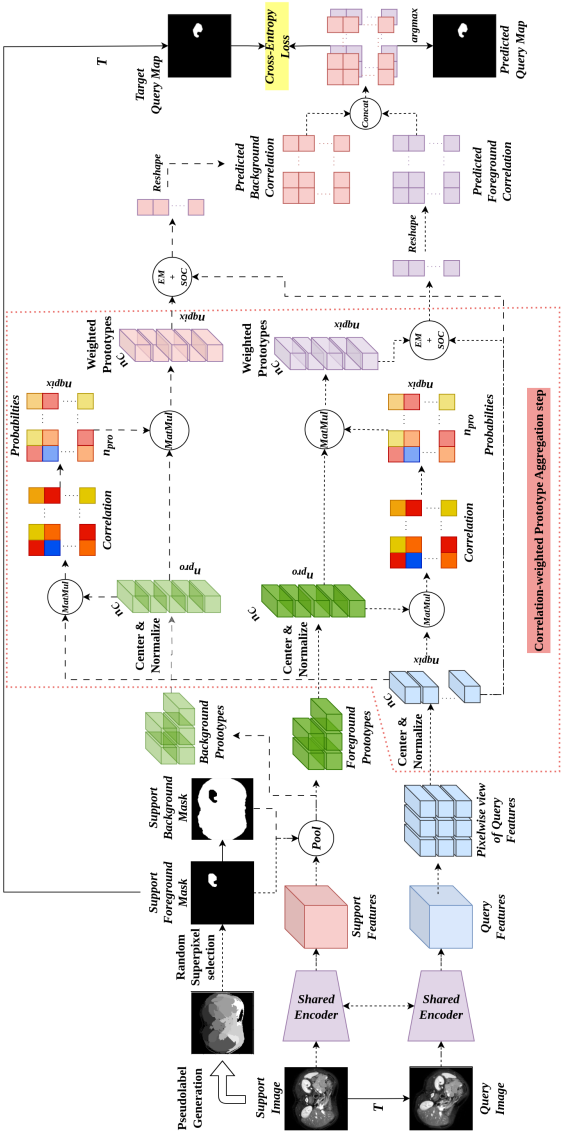


FIGURE 7.1: The figure depicts the entire working principle of the proposed framework. For clarity, we have also indicated the novel proposed correlation-weighted prototype aggregation step using a dotted red bounding box. ***T*** indicates the transformation applied to the support image to generate the query image only in the pre-training stage. ***Pool*** denotes pooling the feature map of the region denoted by the mask. ***MatMul*** denotes Matrix Multiplication. '***EM+SOC***' denotes Element-wise Multiplication and Sum over Channels. ***Concat*** denotes the concatenation operation. n_{pro}, n_{qpix} and n_C denote the number of prototypes, query pixels and channels, respectively. (Best viewed at 200%)

7.4.1 Pretext Task Framework

We propose a self-supervised approach for one-shot learning of segmentation in MR scans. The skeleton of our framework is based on a prototype-based segmentation strategy and consists of the following steps: (1) generation of pseudo-segmentation masks, (2) feature extraction, (3) correlation-weighted prototype aggregation, and (4) mask prediction. In our work, we adopt a dynamic prototype generation approach, one for each query feature map pixel. The dynamic nature of the prototype is the result of the aggregation step using correlation-based probability scores. The final score is obtained by calculating the correlation score of each query pixel to their assigned prototype.

In the downstream segmentation phase, we utilize the slide index information as in Ouyang et al. (2020), to filter out false positives (FP) obtained from other organs or regions on abdominal scans. Furthermore, we also propose to use the quadrant location as a soft spatial information of the respective organ (l) to segment.

7.4.1.1 Generation of Pseudo Segmentation Masks

We follow the same strategy as in Ouyang et al. (2020) for the generation of pseudo-segmentation masks. As stated in Ouyang et al. (2020), superpixel-based segmentation satisfies two properties: for each class, the representations should be clustered to be discriminative under a similarity metric, and the representations should also be invariant across images to ensure robustness. Otherwise, the regions that denote the same class in the support and query images would not be mapped together in the feature space. A superpixel-based clustering strategy ensures that regions with similar pixel features are clustered together. This ensures consistency over each of the pseudo-labels as well.

For every slice in abdominal scans (say I_i), the Felzenszwalb image segmentation algorithm (Felzenszwalb and Huttenlocher, 2004) $\mathcal{F}_{seg}$ is applied to the slice to generate the superpixels, $\mathcal{S}_p = \mathcal{F}_{seg}[I_i])$. During self-supervised training, a superpixel is randomly chosen, $l_s \sim \mathcal{U}[0, |\mathcal{S}_p| - 1]$ and converted to a binary mask to be used as the segmentation mask. A sample of the superpixels obtained is shown in Fig. 7.1.

$$M_s = [\mathcal{S}_p \in \{l_s\}] \tag{7.1}$$

7.4.1.2 Feature Extraction

Each episode $(\mathcal{S}, I_q)$ is passed through the encoder f_θ, which gives us the support and query feature maps, which we denote by $f_\theta(I_s)$ and $f_\theta(I_q)$. In our case, the encoder takes an input of dimension $3 \times 256 \times 256$ and outputs a feature map with dimensions $256 \times 32 \times 32$. We use the *deeplab_v3* version of ResNet101 available from the *torchvision* library. To ensure that the output dimensions match the specifications mentioned above, we used *dilation* in the last two layers of the encoder, similar to Ouyang et al. (2022).

7.4.1.3 Correlation Weighted Prototype Aggregation

The principal component of our proposed framework is the prototype aggregation module, which primarily consists of four steps: 1) Prototype Extraction, 2) Correlation computation, 3) Probability score computation, and 4) Prototype Aggregation. The prototype aggregation steps are done separately for foreground and background.

Prototype Extraction

We do not extract the foreground features by merely doing global average pooling using the support mask M_s. In this case, we follow the steps described in Ouyang et al. (2020), for extracting the foreground and background prototypes. The first step to obtaining the foreground (or background) prototypes is to downsample the segmentation mask to spatial dimension $H \times W$ using an average pooling operation with a window of spatial dimensions 4×4. However, using an average pooling operation may result in values that are not binary (0 or 1). To get a downsampled binary mask, we threshold the interpolated mask. For the foreground, we select a threshold that is 0.8 times the maximum value of the downsampled mask. For the background, we use a threshold that is equal to the mean of the downsampled mask, following Wu et al. (2022a).

$$M_{s(H\times W)} = [AvgPool_{4\times 4}(M_s) > \zeta] \tag{7.2}$$

where $AvgPool$ refers to the Average Pooling operation applied on the binary mask M_s, and ζ is the threshold. However, we find that, for the label sets containing Liver and Spleen, using a threshold of 0.95 for both foreground and background worked better than the aforesaid thresholds (See Table 7.2). Next, the locations $\mathbf{p}$ where the downsampled binarized mask $M_{s(H\times W)}$ is non-zero are processed. The pixels in the support feature map $f_\theta(I_s) \in \mathbb{R}^{D\times H\times W}$, whose locations match those in $\mathbf{p}$, are chosen as prototypes. For the foreground prototypes, we also include the global prototype with the obtained prototypes to avoid an empty set of prototypes resulting from the averaging over the small area of the foreground, following Ouyang et al. (2022).

$$\mathbf{P} = f_\theta(I_s)[M_{s(H\times W)} \in \{1\}] \tag{7.3}$$

Before calculating the cosine similarity between the prototypes $\mathbf{P} \in \mathbb{R}^{D\times N_{pro}}$ and the pixels of the query feature map $f_\theta(I_q)$, we subtract the mean of each of the N_{pro} prototypes along the channel dimension.

$$\mathbf{P}^j = \mathbf{P}^j - \frac{1}{D}\sum_{d=1}^{D} \mathbf{P}^j[d] \tag{7.4}$$

where $\mathbf{P}^j$ is the j^{th} prototype. The steps mentioned above are done for both the foreground and background prototypes. To obtain the foreground prototypes, we simply take M_s as the foreground mask M_s^{FG}, whereas for the background prototypes, we take $M_s^{BG} = 1 - M_s^{FG}$ as the background mask.

Query Features Centering
The same mean subtraction operation is also done for the query pixels in the output feature map, as follows,

$$f_\theta(I_q)_{h',w'} = f_\theta(I_q)_{h',w'} - \frac{1}{D}\sum_{d=1}^{D} f_\theta(I_q)_{h',w'}[d] \tag{7.5}$$

where, $\{h', w'\}$ denotes the location of the pixels in the feature map.

Correlation computation
Having obtained the query feature map $f_\theta(I_q)$ of dimensions $D \times H \times W$ and prototypes $\mathbf{P}$ of dimensions $D \times N_{pro}$, we proceed to compute the cosine similarity between these entities. This results in a correlation matrix or a cosine similarity matrix $\mathcal{M}_c$, which has dimensions $N_{pro} \times H \times W$. This 3D matrix represents the correlation score of all the prototypes obtained in the previous step for each pixel in the query feature map.

$$\mathcal{M}_c(j, h', w') = f_\theta(I_q)(h', w') \odot \mathbf{P}^j \tag{7.6}$$

where $\odot$ indicates the operation of the dot product, $f_\theta(I_q)(h', w')$ denotes the output feature vector at the location (h', w'), and $\mathbf{P}^j$ is the j^{th} prototype. $\mathcal{C}(j, h', w')$ has dimensions $N_{pro} \times H \times W$. The correlation score indicates how similar each prototype is to the query feature map pixels. Intuitively, a prototype $\mathbf{P}^j$ with a higher correlation score with a pixel on the query feature map $f_\theta(I_q)(h', w')$ can be said to be more similar than a prototype with a lower correlation score, in terms of feature similarity. As the feature extractor uses dilation, the receptive field of each pixel in the output feature map has a very large receptive field. Hence, contextual or neighbourhood information is encoded in each foreground prototype $\mathbf{P}^{FG}$. This contextual information will help distinguish the region indicated by the support mask M_s and the other regions.

Probability score computation
The probability of each prototype being similar to a particular query pixel is calculated by taking softmax over the prototypes as follows:

$$\mathcal{M}_{prob}(h', w') = \text{softmax}_{j \in \mathbf{P}}[\mathcal{C}(h', w')/t] \tag{7.7}$$

where $\mathcal{M}_{prob}(h', w')$ denotes the probability of the prototypes $\mathbf{P}$ with respect to the query pixel at the location (h', w'), and τ is a temperature parameter. $\mathcal{M}_{prob}(h', w')$ has dimensions $N_{pro} \times H \times W$.

When calculating the scores for the background prototypes $\mathbf{P}^{BG}$, the background query pixels which are spatially close to a particular background query pixel (say, $f_\theta(I_q)(h', w')$), will yield different correlation scores. This may result in erroneous predictions or increased false positives if we weigh all the prototypes equally. The background prototypes which are spatially farther from $f_\theta(I_q)(h', w')$ region or feature-wise dissimilar will bring the final

score down, thereby increasing false positives. This requires a dynamic prototype that captures contextual information effectively. This is made possible by giving a probabilistic weightage to contextually similar prototypes.

Prototype Aggregation
The weighting scheme necessary for a dynamic and contextual prototype generation is done by aggregating the prototypes based on probability scores obtained from the correlation values between the prototypes and the query pixels. The aggregated dynamic prototype is obtained by a weighted average of all the prototypes using the probabilities obtained in the previous step, as follows:

$$\mathbf{P}_{agg}(h', w') = \sum_{j=1}^{N_{pro}} \mathcal{M}_{prob}(h', w') \cdot \mathbf{P}^j \tag{7.8}$$

where $\mathbf{P}_{agg}(h', w') \in \mathbb{R}^{D\times1\times1}$ denotes the aggregated prototype for the query pixel at location (h', w'), $\mathbf{P}^j \in \mathbb{R}^{D\times1\times1}$ denotes the j-th prototype.

7.4.1.4 Mask Prediction

Computing the Final Score
Once we have the aggregated prototype, we can compute the final scores for each query pixel. The final score is calculated by simply calculating the cosine similarity of the aggregated prototype $\mathbf{P}_{agg}(h', w')$ with the pixel feature of the query in location (h', w'), as follows.

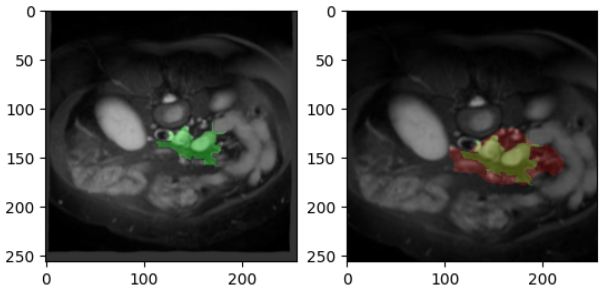
$$c_{FG}(h', w') = \mathbf{P}^{FG}_{agg}(h', w') \odot f_\theta(I_q)(h', w') \tag{7.9}$$

$$c_{BG}(h', w') = \mathbf{P}^{BG}_{agg}(h', w') \odot f_\theta(I_q)(h', w') \tag{7.10}$$

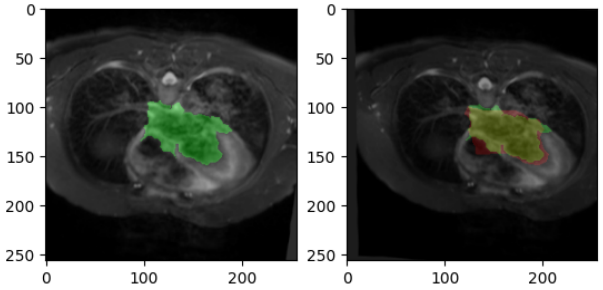
where $c_{FG}(h', w')$ and $c_{BG}(h', w')$ are the scores for the query pixels with respect to the foreground and background prototypes, respectively, and $\mathbf{P}^{FG}_{agg}$ and $\mathbf{P}^{BG}_{agg}$ are the aggregated prototypes for the foreground and background, respectively.

Final Prediction
The final prediction is obtained by choosing the class with the highest probability or the similarity scores for the foreground and background for each query pixel. Thus, the final prediction for each query pixel is obtained as follows:

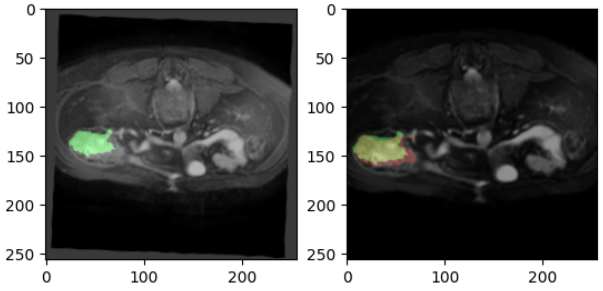
$$\hat{M}_q(h', w') = \underset{\{BG,FG\}}{\text{argmax}}\, \underset{\{BG,FG\}}{\text{softmax}}[c_{BG}, c_{FG}] \tag{7.11}$$



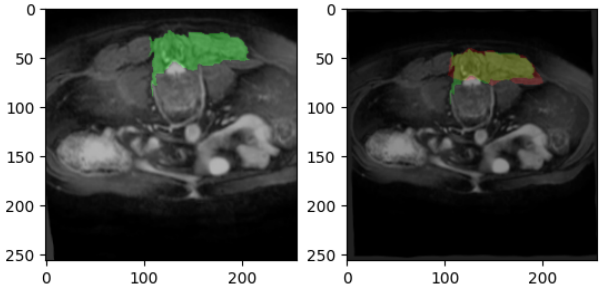
(a) Iteration: 25000



(b) Iteration: 50000



(c) Iteration: 75000



(d) Iteration: 100000

FIGURE 7.2: Predictions in training phase at 25K, 50K, 75K, 100K iterations. The left image in Figs. (a)-(d) is the support image I_s and the support mask is denoted in *green*. The right image in Figs. (a)-(d) is the query image. The ground truth is denoted by *green* and the predicted mask is indicated by *red*. (Use 200% zoom for better visibility)

where $\hat{M}_q(h', w')$ is the predicted query mask. A few examples of the query predictions and the associated ground truth from the training stage are shown in Fig. 7.2.

7.4.1.5 Training Pipeline

In each iteration t, we take an episode $((I_s, M_s), I_q)$ as input, and the model predicts $\hat{M}_q$ as output. The query image I_q is obtained by applying geometric and intensity transformations on the support image I_s, that is, $I_q = \mathcal{T}_{geo}(\mathcal{T}_{int}(I_s))$. The pseudo-ground truth is obtained by only applying the geometric transformation $\mathcal{T}_{geo}$ on the support mask M_s.

Geometric transformations $\mathcal{T}_{geo}$ consist of affine and elastic transformations to emulate the changing shapes of the class labels in the downstream task. Intensity transformations $\mathcal{T}_{int}$ consist of gamma transformation to account for the varying intensity of the pixels between scans of different patients.

Since the encoder f_θ output is of spatial dimension 32×32, we interpolate the final prediction to 256×256 before calculating the loss using bilinear interpolation. To optimize the model parameters, we minimize the cross-entropy loss $\mathcal{L}^t_{ssl}$ for all the query pixels.

$$\mathcal{L}^t_{ssl}(\theta) = - \mathop{\mathbb{E}}_{h',w'} \left[\lambda_{h',w'} \mathcal{T}_{geo}(M^t_s)(h', w') \log \left(\hat{M}^t_q(h', w') \right) \right] \tag{7.12}$$

where $\hat{M}^t_q(h', w')$ is the predicted output of $\mathcal{T}_{geo}(M^t_s)(h', w')$ taking $I_q = \mathcal{T}_{geo}(\mathcal{T}_{int}(I_s))$ as query. (h', w') denotes the location in the predicted query mask or pseudo-ground-truth mask. $\lambda_{h',w'}$ denotes the class weight applied during training.

Similar to Ouyang et al. (2020), we also adopt the Cyclic Consistency Regularization (CCR) following Wang et al. (2019). To implement CCR, we interchange query and support images. For the support mask, we use the predicted query output $\hat{M}_q$ as the foreground mask, and the support mask M_s in the forward iteration is used as the pseudo-ground-truth. The episode in the CCR step consists of $((I_q, \hat{M}_q), I_s)$. In the CCR step, we use an initial threshold of 0.95 in the prototype extraction step to filter out noisy predictions, following which the aforementioned steps are followed. Otherwise, we see a drop in performance by about 2% in dice score. The CCR loss is represented as

$$\mathcal{L}^t_{reg}(\theta) = - \mathop{\mathbb{E}}_{h',w'} \left[M^t_s(h', w') \ln \left(\hat{M}^t_s(h', w') \right) \right] \tag{7.13}$$

where $\hat{M}^t_s(h', w')$ is the predicted output of $\hat{M}^t_q(h', w')$ taking I_s as a query.

Hence, the total loss is as follows,

$$\mathcal{L}^t = \mathcal{L}^t_{ssl} + \mathcal{L}^t_{reg} \tag{7.14}$$

To handle the imbalance, we set the class weights at 0.05 for the background pixels or the class label 0, and a weight of 1.0 for the foreground pixels or the class label 1.

Furthermore, it is to be noted that during training, we divide the class labels in abdominal MR into two parts, namely, upper abdomen consisting of *right kidney* and *left kidney*, and lower abdomen consisting of *liver* and *spleen*. When training on slices from the upper abdomen, we do not include slices containing lower abdomen classes and vice versa.

7.4.2 Downstream Task Framework

The primary objective of this chapter was to propose a pretext task which is similar to the downstream task to minimize the task disparity and learn representations which are better aligned to the downstream task. Hence, all the steps in the downstream task are the same as the ones in the pretext task, except for the generation of pseudo masks. The steps involved in the downstream tasks are (1) Feature Extraction, (2) Correlation weighted prototype aggregation, and (3) Mask Prediction.

7.4.2.1 Validation without Fine-tuning

Following Ouyang et al. (2020, 2022), we evaluate our model on a validation split, without further fine-tuning on the whole dataset. Although we don't fine-tune the model, we use the class label information from the dataset. For a class label l, we only take the slices $[u_{min}, u_{max}]$ in which l is present for the final predictions.

7.4.2.2 One-Shot Segmentation

The evaluation strategy follows a one-shot segmentation task. Among the scans in the validation split, the last scan (if arranged in order) is selected as the support scan. The sequential range of slices $[u_{min}, u_{max}]$ in both support and query scans is divided into 3 parts following the evaluation strategy followed in Guha Roy et al. (2020). From the three support scan splits, the middle slice is selected as the support image for the whole of the corresponding query part. The evaluation step then follows the flow of the one-shot segmentation task, that is, a tuple $((I_s^u, M_s^u), I_q^{u,i})$ is fed to the pre-trained model f_θ as input to obtain the predicted query mask $\hat{M}_q^{u,i}$, where u is the part in which the slices belong and $i \in [u_{min}^q, u_{max}^q]$ is the index of the slices of query scan in which the class label l is present.

7.4.2.3 Quadrant Masking Scheme

During training, the classes in the abdominal MR dataset are split into two parts, the upper abdomen (right and left kidneys) and the lower abdomen (liver and spleen). During validation, we divide each slice into quadrants. For each class label l, we identify which quadrants are occupied by it. The final predictions obtained from the one-shot segmentation step are masked such that the predictions from the quadrants in which the class label l is present, are considered for the final metric calculation. For example, the class *right*

kidney is present in the left half of an MR slice. Hence, we mask the right half of each slice while making the final prediction.

The quadrant masking scheme uses the quadrant information as a piece of soft prior information about the location of the target organ. To the best of our knowledge, no work has employed this scheme before. To fully understand the role of this quadrant masking scheme, we conduct an ablation study in Sec. 7.5.4.2, where we observe the significant effects of the soft prior knowledge in boosting the segmentation performance.

7.4.2.4 Validation Metric

To measure the performance of the proposed model, we use the Dice score as a metric, as is usually done in the medical image segmentation literature (Ouyang et al., 2020, 2022; Wang et al., 2019).

7.5 Experiments Details, Results and Analysis

In this section, we discuss the details of the datasets used, and the implementation details, followed by the quantitative and qualitative results, and ablation studies.

7.5.1 Datasets

To demonstrate the effectiveness of the proposed approach, we used Magnetic Resonance (MR) scans in our work. The data used in our work contained scans from different people with varying medical conditions as observed from inspecting the data.

For the Magnetic Resonance (MR) scans, we used the Combined Healthy Abdominal Organ Segmentation (CHAOS) Challenge (Task 5) from ISBI 2019 (Kavur et al., 2021). This dataset contains 20 3D T2-SPIR MRI scans.

For the experiments, we used a five-fold cross-validation setting, that is, in an experimental run, one-fifth of the dataset (a fold) is used as a validation set while the rest is treated as a training set.

7.5.2 Implementation Details

Training and evaluation were implemented using PyTorch. The training was done on a 24GB NVIDIA A5000 GPU. The average training time for each training run consisting of 100K iterations was about 4.5 hours. We used a batch size of 1. The initial learning rate of the SGD optimizer was set to $1e-3$ and decayed at 0.95 per 1K iterations.

TABLE 7.1: DICE score on Abdominal MR (CHAOS) Dataset. Reported Values are with Single Support Scan. 'Sup.' stands for Supervised. ✓in the 'Sup' column indicates the corresponding method is a supervised one, and × indicates otherwise.

Method	Sup.	RK	LK	Liver	Spleen	Mean
SE-Net (Guha Roy et al., 2020)	✓	61.32	62.11	27.43	51.80	50.66
Vanilla PANet (Wang et al., 2019)	✓	38.64	53.45	42.26	50.90	46.33
ALPNet (Ouyang et al., 2020)	✓	58.99	53.21	37.32	52.18	50.43
SSL-PANet (Ouyang et al., 2020)	×	47.95	47.71	64.99	58.73	54.85
SSL-ALPNet (Ouyang et al., 2020)	×	78.39	73.63	73.05	67.02	73.03
CRAPNet (Ding et al., 2023)	×	**82.77**	<u>74.66</u>	<u>73.82</u>	<u>70.82</u>	<u>75.52</u>
CoWPro (Proposed)	×	<u>80.45</u>	**75.30**	**75.77**	**71.51**	**75.56**

7.5.3 Comparative Results and Analysis

In this section, we present both quantitative and qualitative performance analyses of the proposed framework. Quantitative analysis refers to the comparison of the performance in terms of dice score, whereas qualitative analysis refers to displaying predicted masks with respect to the ground truth and inferring the quality of the segmentation.

7.5.3.1 Quantitative Performance Analysis

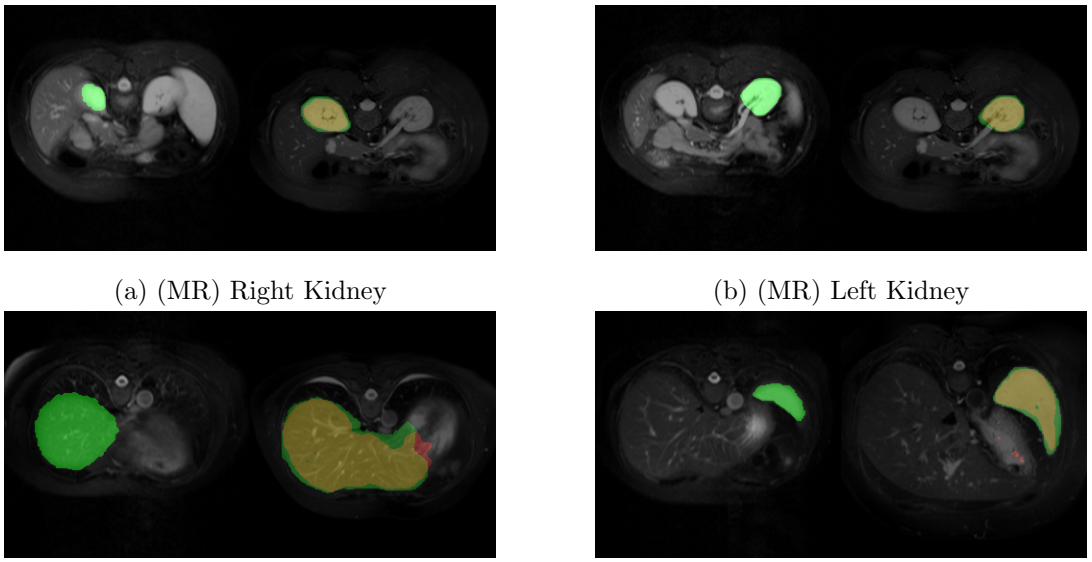
In this section, we present the results obtained by the proposed model on the MR datasets CHAOS (Kavur et al., 2021). From Table 7.1, we can see that the proposed framework outperforms ALPNet (Ouyang et al., 2020) and also outperforms several current state-of-the-art methods in several classes. The bold font and the underlined text indicate the best and the second-best performance, respectively. The proposed algorithm outperforms CRAPNet on the CHAOS dataset without any further fine-tuning of hyperparameters. On the CHAOS dataset, the proposed framework outperforms the SOTA method on the Left kidney, Liver and Spleen. This shows that the dynamic prototype aggregation technique improves the representation learning and generalizability of the model over the single prototype-based SSL-ALPNet framework.

7.5.3.2 Qualitative Performance Analysis

The qualitative performance of the proposed model can be seen from the predictions presented in Fig. 7.3. We can see the predictions for different organs on two different modalities compared to the ground truth. We can see that the model produces segmentation results close to the ground truth.

7.5.4 Ablation Studies

In this section, we study the effects of different parameters on the performance of the proposed framework in the downstream task of one-shot segmentation.



(a) (MR) Right Kidney (b) (MR) Left Kidney

(c) (MR) Liver (d) (MR) Spleen

FIGURE 7.3: Figure showing the predictions obtained for 4 organs, Right Kidney, Left Kidney, Liver, and Spleen for MR (CHAOS dataset) scans. (green) Ground Truth, (red) Prediction, (yellow) Ground Truth and Prediction overlap. (Use 300% zoom for better visibility)

7.5.4.1 Effect of Fixed vs. Dynamic Threshold

The effect of threshold on the performance of the proposed framework can be seen in Table 7.2, where we see the use of two different types of threshold, dynamic and fixed, for different sets of labels. In the CHAOS dataset, the use of a dynamic thresholding scheme similar to the one used in Wu et al. (2022a) works better for the labels *right kidney* and *left kidney*, whereas a fixed threshold of 0.95, similar to Ouyang et al. (2022) works better for the other two labels. The dynamic thresholding scheme assigns a threshold of 0.8 times the max value of the downsampled foreground mask, whereas for the background it uses the mean value of the downsampled background mask. The dynamic thresholding scheme allows the model to adapt to the local pixel intensities and infuses local spatial information in the process. However, for large organs like the spleen and liver, the effect of dynamic thresholding is not positive.

TABLE 7.2: Dice score obtained on abdominal MR dataset CHAOS using different thresholding schemes without Quadrant masking scheme.

Threshold	Abdominal MR			
	R. Kidney	L. Kidney	Liver	Spleen
Fixed	79.06	72.24	**75.04**	**71.19**
Dynamic	**79.54**	**74.59**	73.87	66.33

TABLE 7.3: Dice score obtained on abdominal MR dataset (CHAOS) with and without quadrant masking scheme with a fixed threshold.

Quadrant Masking	Abdominal MR			
	R. Kidney	L. Kidney	Liver	Spleen
Yes	**79.66**	**75.30**	**75.77**	**71.51**
No	79.06	72.24	75.04	71.19

7.5.4.2 Effect of Quadrant Masking

In Tab. 7.3, we observe the effect of the quadrant masking scheme on the segmentation performance on the MR dataset. Barring a few exceptions, the quadrant masking scheme has improved the dice score for all the organs. The primary reason behind the slight drop in performance can be attributed to the hard quadrant boundary assigned to the corresponding organs.

7.5.4.3 Effect of Number of Aggregated Prototypes

The correlation-weighted prototype-aggregation step aims to incorporate the information from all the prototypes to prevent loss of information, as is generally the case in prototype-based methods. However, one may argue that using all the prototypes may induce an unintended negative effect from anatomically different but semantically similar regions, consequently causing a degradation in performance. In this ablation study, we take the Top-100%, 50%, 10%, 5%, and 2% most similar prototypes to predict foreground or background regions. This study also establishes the efficiency of our model in encoding contextual information, which is evident from the insignificant variation in the dice scores with the decrease in the number of prototypes. In Tab. 7.4, we show the dice scores for varying numbers of prototypes in the inference stage for the left and right kidney over all the folds.

TABLE 7.4: Dice score for 2×2 and 4×4 averaging window without Quadrant Masking.

Averaging Window		2×2		4×4	
Organ		RK	LK	RK	LK
Percentage of prototypes	100%	**79.99**	76.22	80.77	72.87
	50%	79.89	**76.24**	80.72	72.94
	10%	79.49	75.22	80.94	73.81
	5%	78.16	73.46	**81.08**	74.43
	2%	75.69	70.42	80.68	**74.86**

7.5.4.4 Effect of Averaging window

In Tab. 7.4, we observe the effect of changing the averaging window in the prototype generation step. We chose the left and right kidneys from the abdominal MR dataset to study the effect of the averaging window, as the two kidneys are almost similar in shape

and size but vary only in the spatial context. We observe that using the same averaging window as in training, that is, an averaging window of 4×4, the performance of the proposed framework is better for the right kidney, than using an averaging window of 2×2. On the contrary, using an averaging window of 2×2 yields better performance than using an averaging window of 4×4 on the left kidney. We believe that this discrepancy in the trend is primarily due to the different spatial contexts of the two organs.

7.6 Conclusion

In this work, we have presented a prototype-based framework for self-supervised one-shot learning of medical image segmentation tasks. Instead of taking a pre-training representation learning approach, we take a task-learning-based approach. To deal with the issue of variations in the background information between the support and query images, we propose a correlation-based weighting scheme to aggregate the support prototypes according to how related the prototypes are to the query image. Therefore, each query feature map pixel has a customized prototype. The score for foreground or background is obtained by calculating the cosine similarity of the query feature map pixels with their corresponding prototype. The primary objective of constructing a prototype for each query feature map pixel is to reduce false positives in the predictions by weighing down the contribution of dissimilar prototypes in the final prediction. Despite the limitations of the proposed method, we can see that the proposed method outperforms most of the contemporary self-supervised segmentation methods.

Chapter 8

Conclusion and Future Directions

This chapter provides a concise summary of the primary contributions of the research work done under this thesis. Additionally, it outlines the future scope of the work, including potential extensions and applications.

8.1 Introduction

In this thesis, we primarily investigate the paradigm of Self-supervised learning mainly for medical image-based tasks (we also utilise natural images in some of our proposed frameworks to benchmark and learn representations in some cases). Self-supervised learning is a sub-category of unsupervised learning. This learning paradigm involves designing a pretext task, which aids the learning of representations that are better suited for the downstream task than transferred weights pre-trained on datasets with data distribution different from the dataset used in the downstream task. The pretext task can be of different types, context-based, instance discrimination-based, etc.

The quality of the representations depends on the pretext task. Apart from challenges like sensitivity to data augmentation and the requirement of heavy computational resources, the SSL framework faces several other challenges. For different types of frameworks, these challenges are different. For the context-based frameworks with pretext tasks like rotation prediction, jigsaw puzzle solving, etc, the representations learnt are not context invariant. This results in the representations not being of the quality necessary to yield satisfactory performance in the downstream task. We aimed to propose a novel framework that learns context-invariant representations and improves representation learning on small-scale medical image datasets.

Furthermore, before investigating deeper into self-supervised learning, we need to investigate if self-supervised pre-training affects downstream performance. For this purpose, a novel binary contrastive learning framework is proposed. Additionally, we attempt to improve this binary contrastive learning framework by carefully striking a balance between alignment and uniformity.

In the self-supervised learning paradigm, the concept of convergence is not well understood and there are only a few pieces of work on this. Hence, one of our objectives in this thesis is to study the phenomenon of convergence in SSL frameworks both mathematically and empirically and find the necessary and sufficient conditions under which convergence is guaranteed.

For instance discrimination-based self-supervised contrastive learning frameworks, the temperature plays a significant role in controlling the uniformity and alignment in the feature space by influencing the attraction and repulsion between the positive and negative pair samples. The temperature hyper-parameter also influences the contribution of the hard and false negative pairs in the representation learning process. However, there are only a limited number of studies on this. Starting from a single assumption, we aim to propose a temperature function which follows a few ideal tenets for a temperature function to maximize the quality of representations.

Another intriguing aspect of current context-based and instance discrimination-based frameworks is that the pretext and downstream tasks have different objectives. As already discussed, the quality of representations depends on the pretext task. Also, the transferability of the representations depends on how much the pretext and downstream tasks differ. Keeping the objective of the pretext and downstream task the same, the representations learnt in the pretext task are better suited to the downstream task. Therefore, our objective is to learn better transferable representations by devising the same one-shot segmentation framework in both the pretext and downstream tasks.

Hence, we propose several novel frameworks to deal with the above mentioned gaps in the self-supervised learning paradigm.

8.2 Summary of Contributions

Based on the above mentioned research gaps, this thesis aims to propose several novel frameworks, primarily under the subcategory of context-based frameworks and instance discrimination-based contrastive frameworks.

First and foremost, a detailed overview of the self-supervised learning landscape is presented in Chapter 2. In this chapter, we explore the various types of self-supervised frameworks, categorizing them by their underlying principles, and critically analyze the foundational approaches. We also review studies that leverage SSL frameworks for learning representations from medical image data, organizing them by imaging modality. Additionally, this chapter highlights the strengths and limitations of these approaches.

To learn context-invariant representations in context-based frameworks, in Chapter 3, we primarily use the jigsaw puzzle-solving objective to learn representations. To prevent learning of redundant and low-level representations, a novel architecture which uses semi-parallel convolutional blocks to effectively decouple the context dependence between input patches of the branches has been proposed. Furthermore, to boost generalization, we concatenate the output from the branches along the feature dimensions. We also investigate the effect of class imbalance on pretext training, and consequently on the downstream task

performance. Another finding of this chapter is that bigger self-supervised models perform better on downstream tasks.

In Chapter 4, our contributions are two-fold. Firstly, we propose a novel contrastive learning framework termed the Binary Contrastive learning framework based on noise contrastive estimation. Secondly, we investigate the effectiveness of self-supervised pre-training over ImageNet pre-trained representations on medical visual data.

In Chapter 5 too, there is more than one contribution. Firstly, the imbalance inherent in the formulation of the Binary Contrastive learning framework proposed in Chapter 4 was dealt with by reformulating the Binary Contrastive learning maximum likelihood objective in a variational approach. By controlling the influence of the sample in a negative pair, we intended to improve representation learning by finding a better trade-off between the uniformity and alignment metric. Furthermore, this variational formulation was improved by discarding the positive pair repulsion term and taking an upper bound of the negative pair repulsion term. The comparative performance demonstrates the superiority of the proposed method on various downstream tasks. Additionally, we conduct a mathematical and empirical analysis to investigate whether SSL frameworks converge to local or global minima. Through evaluation of eigenspectrum, it was found that SSL frameworks converge only under certain conditions and under a long duration of training. We also observe the influence of temperature on the different versions of the binary contrastive learning framework.

In Chapter 6, we delve deeper into this aspect of the temperature hyper-parameter on the self-supervised representation learning process. The temperature hyper-parameter influences the uniformity-alignment trade-off and also controls the repulsion of the hard false negative pair samples. From a simple intuitive assumption, we obtain a first-order differential equation describing the variation of temperature with the cosine similarity of a sample pair. Solving this first-order ordinary differential equation we obtain the temperature function which prevents samples in false negative pairs from drifting too far and also amplifies the repulsion between true negative pairs, without affecting convergence. The findings and claims are also supported by empirical results which show that the proposed framework can outperform state-of-the-art SSL frameworks by boosting the performance of the weakest baseline on benchmark datasets.

The performance in the downstream task depends on the quality and transferability of representations learnt in the pretext task. The representations learnt in a pretext task are not always suited to the downstream task as the objective differs. To ensure task similarity between the pretext and downstream phase, in Chapter 7, we intend to minimize the task disparity between the pretext and downstream task by using a self-supervised few-shot task for representation learning as the pretext task trained for the task of one-shot segmentation in the downstream stage. For the pretext task, we use pseudo-masks that were generated using the Fenzelschwalb segmentation algorithm and a correlation-weighted prototype aggregation-based approach is used to predict the similarity with the foreground and background prototype features for segmentation. The proposed approach performs better or at par with the contemporary self-supervised segmentation frameworks on abdominal multi-organ segmentation tasks.

In short, the frameworks proposed in this thesis cover a wide range of principles and aim to solve various challenges in SSL. Not only that, the proposed frameworks have also succeeded in outperforming the contemporary state-of-the-art methods in most cases. The novel ideas explored in this thesis also pave the way for future research in this domain.

8.3 Limitations

Discussing the limitations of a thesis is important in many aspects. It allows the readers to understand the boundaries of the study, and the context in which the results should be interpreted. Thus, it provides a lot of transparency. It also increases the credibility of the research and the thinking capability of the researcher as well. It also helps in building trust and also spreads positive indications about the reliability of the work.

This study primarily focuses on three types of frameworks, context-based frameworks, instance discrimination-based contrastive learning frameworks and a few-shot-based SSL frameworks. In the context-based framework using the jigsaw puzzle-solving strategy discussed in Chapter 3, we only experimented on two combinations of jigsaw configurations. This limits the study on the effect of the complexity of the pretext task on the representation learning process, and consequently the downstream performance. Furthermore, we only divide each input into 3×3 parts (which gives us 9! permutations), as increasing the number of divisions increases the number of permutations (for instance, 4×4 parts will lead to 16! permutations). Thus, sampling a fixed number of jigsaw configurations from an increased number (16!) of configurations would take an exponentially large amount of time.

In the instance discrimination-based task in Chapter 5 and Chapter 6, we conducted all experiments on a maximum available 32GB GPU. Hence, it was not possible to accommodate more than a batch size of 256 even with automatic mixed precision training. This restricted our capability to run experiments with a large batch size and also larger models. Furthermore, an experiment on ImageNet (Deng et al., 2009) dataset for 100 epochs of self-supervised pre-training took approximately 7 days. Thus, training our proposed frameworks on the Imagenet dataset for 800 or 1000 epochs proved difficult, as is often the norm in current SSL research. Consequently, we were also unable to provide transfer learning performances on tasks like object detection, object segmentation, etc.

8.4 Future Scopes of Research

The research presented in this thesis can be extended further for the progress of the self-supervised learning domain. Here, we list some directions that may be pursued in the future.

- **Combining Contrastive and Non-Contrastive**: While contrastive and non-contrastive learning frameworks have made progress as separate lines of research within the canopy of SSL, we believe combining the positive aspects of contrastive

learning like prevention of complete collapse with that of non-contrastive learning like alignment maximization can boost representation learning.

- **Combining Instance Contrastive and Implicit Variance Regularization**: Similar to the above point, instance-based contrastive learning and Implicit variance regularization principle prevent two different types of collapse. Hence, combining these two types of principles can also improve representation learning.

- **FP8 Training**: As the support for 8-bit floating point training (Micikevicius et al., 2022; Kuzmin et al., 2022; Agrawal et al., 2024) gets distributed for more GPUs, the representation learning paradigm can be improved by increasing the batch size and faster training. This will lead to faster experimentation and subsequently faster innovations. While FP8 has already been applied to finetune resource-hungry frameworks like diffusion models, large language models or large multi-modal models like GPT-4o (Achiam et al., 2023) can also be trained using self-supervised learning and fine-tuned on low-resource environments.

- **Generative SSL**: Recently, diffusion models have gained much popularity and progress. Self-supervised representation learning can be described as learning the data distribution over a low-dimensional manifold. Hence, incorporating generative modelling within SSL frameworks like unconditional GANs or diffusion models can boost the learning of representations.

Appendix A

Understanding Convergence on Non-Convex Functions with Polyak-Lojasiewucz Inequality

A.1 Convergence on Non-Convex Functions

A.1.1 Polyak-Lojasiewicz Inequality

From Karimi et al. (2016), we can state, for an unconstrained optimization problem,

$$\text{argmin}_{x \in \mathbb{R}^d} f(x) \tag{A.1}$$

where f is a function with L-Lipschitz continuous gradient, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2 \; \forall \; x, y \tag{A.2}$$

If f belongs to the class of C^2 functions, the eigenvalues of $\nabla^2 f(x)$ are bounded above by L, which is called the Lipschitz constant. We also assume that the solution set $\mathcal{X}^* \neq \varnothing$ and f^* is the optimal function value. The Polyak-Lojasiewicz inequality is satisfied if for $\mu > 0$,

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*), \; \forall \; x \tag{A.3}$$

Applying gradient descent with step size $\frac{1}{L}$,

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) \tag{A.4}$$

From Eq. A.2, we get,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\ &\leq f(x_k) - \frac{\mu}{L}\left(f(x_k) - f^*\right) \end{aligned} \tag{A.5}$$

Subtracting f^* from both sides of Eq. A.5, we get,

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)\left(f(x_k) - f^*\right) \tag{A.6}$$

Applying Eqn. A.6 recursively, we get, the global linear convergence rate as follows,

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_0) - f^*\right) \tag{A.7}$$

A.1.2 Convergence of SGD on Non-Convex Functions

To calculate the rate of convergence of SGD on non-convex functions, we follow the derivation steps followed in Orabona (2020). To see how SGD evolves over time, we take $x = w_t, y = w_{t+1}$ in Eqn. A.2, which yields,

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2}\|w_t - w_{t+1}\|^2 \tag{A.8}$$

Now, we assume that for the samples indicated by the indices ξ, we have an oracle that gives us the gradient $g(w, \xi) \in \mathbb{R}^d$ at the point w, where ξ is a random index of a training sample used to calculate the training loss. We also assume that the variance of the stochastic gradient is bounded by as $\mathbb{E}_\xi\left[\|\nabla f(w) - g(w, \xi)\|_2^2\right] \leq \sigma^2 < \infty$ for all $w \in dom \nabla f(w)$.

For SGD, the parameter update proceeds as $w_{t+1} = w_t - \eta_t g(w_t, \xi_t)$, where η_t is the learning rate at time step t. Now, going back to Eqn. A.8, and putting $w_{t+1} - w_t = -\eta_t g(w_t, \xi_t)$, we get,

$$f(w_{t+1}) \leq f(w_t) - \eta_t \langle \nabla f(w_t), g(w_t, \xi_t) \rangle + \eta_t^2 \frac{L}{2}\|g(w_t, \xi_t)\|_2^2 \tag{A.9}$$

Taking expectation with respect to ξ_t, keeping w_t constant, we get

$$
\begin{aligned}
&\mathbb{E}_{\xi_t}\left[f(w_{t+1})\right] \leq \mathbb{E}_{\xi_t}\left[f(w_t)\right] - \eta_t \mathbb{E}_{\xi_t}\left[\langle \nabla f(w_t), g(w_t, \xi_t)\rangle\right] + \eta_t^2 \frac{L}{2} \mathbb{E}_{\xi_t}\left[\|g(w_t, \xi_t)\|_2^2\right] \\
\implies\ & f(w_{t+1}) = f(w_t) - \eta_t \|\nabla f(w_t)\|_2^2 + \eta_t^2 \frac{L}{2} \mathbb{E}_{\xi_t}\left[\|g(w_t, \xi_t)\|_2^2\right] \\
&\qquad = f(w_t) - \eta_t \|\nabla f(w_t)\|_2^2 \\
&\qquad + \eta_t^2 \frac{L}{2} \mathbb{E}_{\xi_t}\left[\|\nabla f(w_t) + g(w_t, \xi_t) - \nabla f(w_t)\|_2^2\right] \\
&\qquad = f(w_t) - \eta_t \|\nabla f(w_t)\|_2^2 + \eta_t^2 \frac{L}{2} \left(\mathbb{E}_{\xi_t}\left[\|g(w_t, \xi_t) - \nabla f(w_t)\|_2^2\right] + \|\nabla f(w_t)\|_2^2\right) \\
&\qquad = f(w_t) - \left(\eta_t - \frac{\eta_t^2 L}{2}\right) \|\nabla f(w_t)\|_2^2 + \eta_t^2 \frac{L}{2} \mathbb{E}_{\xi_t}\left[\|g(w_t, \xi_t) - \nabla f(w_t)\|_2^2\right] \\
&\qquad \leq f(w_t) - \left(\eta_t - \frac{\eta_t^2 L}{2}\right) \|\nabla f(w_t)\|_2^2 + \eta_t^2 \frac{L}{2} \sigma^2
\end{aligned}
\tag{A.10}
$$

In the last line of the above equation, we have used the fact that the variance of the stochastic gradient is bounded above by σ^2. Taking the total expectation and reordering terms, we get,

$$
\begin{aligned}
&\sum_{t=1}^{T} \left(\eta_t - \frac{\eta_t^2 L}{2}\right) \mathbb{E}_t\left[\|\nabla f(w_t)\|_2^2\right] \\
&\leq \sum_{t=1}^{T} \left(\mathbb{E}_t\left[f(w_t)\right] - \mathbb{E}_t\left[f(w_{t+1})\right]\right) + \frac{\sigma^2 L}{2} \sum_{t=1}^{T} \eta_t^2 \\
&\leq \mathbb{E}_t\left[f(x_1)\right] - \mathbb{E}_t\left[f(w_{T+1})\right] + \frac{\sigma^2 L}{2} \sum_{t=1}^{T} \eta_t^2 \\
&\leq f(x_1) - f^* + \frac{\sigma^2 L}{2} \sum_{t=1}^{T} \eta_t^2
\end{aligned}
\tag{A.11}
$$

Optimization with Constant Step Size

The optimization process continues as long as $\eta_t < \frac{1}{L}$, where L is the Lipschitz constant. The left-hand side of the above equation will be maximized for $\eta_t = \frac{1}{L} = \eta$. Hence, putting that value, we have $\eta_t - \frac{\eta_t^2 L}{2} = \eta_t - \frac{\eta_t}{2} = \frac{\eta_t}{2}$. Putting this expression in Eqn. A.11, we get

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{\eta_t}{2} \mathbb{E}_t\left[\|\nabla f(w_t)\|_2^2\right] \leq f(x_1) - f^* + \frac{\sigma^2 L}{2} T \eta_t^2 \\
\implies\ & \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_t\left[\|\nabla f(w_t)\|_2^2\right] \leq \frac{2}{\eta_t T} (f(x_1) - f^*) + \sigma^2 L \eta_t \\
&\qquad = \frac{2L}{T} (f(x_1) - f^*) + \sigma^2
\end{aligned}
\tag{A.12}
$$

We get an almost convergence result, as the average of the norm of the gradients goes to zero at $\mathcal{O}(\frac{1}{T})$. This means that we can expect the algorithm to make fast progress at the beginning of the optimization and then slowly converge once the number of iterations becomes big enough compared to the variance of the stochastic gradients. In case the noise on the gradients is zero, SGD becomes simply gradient descent and it will converge at a rate of $\mathcal{O}(\frac{1}{T})$.

Optimization with Time-Varying Step Size

Let us consider again Eqn. A.11, but with a time-varying learning rate,

$$\sum_{t=1}^{\infty} \eta_t = \infty \text{ and } \sum_{t=1}^{\infty} \eta_t^2 < \infty \tag{A.13}$$

The above conditions ensure that $\eta_t \to 0$ as $t \to \infty$ (Bottou et al., 2018).

With such a choice, we get,

$$\begin{aligned} &\sum_{t=1}^{T} \left(\eta_t - \frac{\eta^2 L}{2} \right) \mathbb{E}_t \left[\|\nabla f(w_t)\|_2^2 \right] \\ &\leq f(w_1) - f^* + \frac{\sigma^2 L}{2} \sum_{t=1}^{T} \eta_t^2 < \infty \end{aligned} \tag{A.14}$$

Now, $\sum_{t=1}^{T} \eta_t^2 < \infty \implies \eta_T \to 0$. So, there exists T_L such that $\eta_t - \frac{\eta_t^2 L}{2} \geq \frac{\eta_t}{2}$ for all $t \geq T_L$. Hence,

$$\sum_{t=T_L}^{\infty} \eta_t \mathbb{E}_t \left[\|\nabla f(w_t)\|_2^2 \right] < \infty \tag{A.15}$$

This implies that $\sum_{t=T_L}^{\infty} \eta_t \|\nabla f(w_t)\|_2^2 < \infty$ with probability 1. From this last inequality and the condition $\sum_{t=1}^{\infty} \eta_t = \infty$, we can derive that $lim\ inf_{t\to\infty} \|\nabla f(w_t)\|_2 = 0$.

Unfortunately, it seems that we proved something weaker than we wanted to. In words, the *lim inf* result says that there exists a subsequence of w_t that has a gradient converging to zero.

Lemma A.1.1. *Let $(b_t)_{t\geq 1}, (\eta_t)_{t\geq 1}$ be two non-negative sequences and $(a_t)_{t\geq 1}$ a sequence of vectors in a vector space X. Let $p \geq 1$ and assume $\sum_{t=1}^{\infty} \eta_t b_t^p < \infty$ and $\sum_{t=1}^{\infty} \eta_t = \infty$. Assume also that there exists $L \geq 0$ such that $|b_{t+\tau} - b_t| \leq L(\sum_{i=t}^{t+\tau-1} \eta_i b_i + \|\sum_{i=t}^{t+\tau-1} \eta_i a_i\|)$, where a_t is such that $\|\sum_{i=1}^{\infty} \eta_t a_t\| < \infty$. Then, b_t converges to 0. [Lemma A.5 in Mairal (2013), Extension of Proposition 2 in Alber et al. (1998)]*

Using the above Lemma on $b_t = \|\nabla f(w_t)\|$, we observe that by the L-smoothness of f, we have,

$$\begin{aligned}
&\|\nabla f(w_{t+\tau})\| - \|\nabla f(x_t)\| \leq \|\nabla f(w_{t+\tau}) - \nabla f(w_t)\| \\
&\leq L\|w_{t+\tau} - w_t\| = L\|\sum_{i=t}^{t+\tau-1} \eta_i g(x_i, \xi_i)\| \\
&= L\|\sum_{i=t}^{t+\tau-1} \eta_i \left(\nabla f(x_i) + g(x_i, \xi_i) - \nabla f(x_i)\right)\| \\
&\leq L \sum_{i=t}^{t+\tau-1} \eta_i \|\nabla f(x_i)\| + L\|\sum_{i=t}^{t+\tau-1} \eta_i \left(g(x_i, \xi_i) - \nabla f(x_i)\right)\|
\end{aligned} \tag{A.16}$$

The assumptions and the reasoning above imply that, with probability 1, $\sum_{t=1}^{\infty} \eta_t \|\nabla f(w_t)\| < \infty$. This also suggest to set $a_t = g(x_i, \xi_i) - \nabla f(x_i)$. Also, we have, with probability 1, $\|\sum_{t=1}^{\infty} \eta_t a_t\| < \infty$, because $\sum_{t=1}^{T} \eta_t a_t$ for $T = 1, 2, ...$ is a martingale, i.e., the conditional expectation of the next value in the sequence is equal to the present value, regardless of all prior values. The variance is also bounded by $\sigma^2 \sum_{t=1}^{\infty} \eta_t^2 < \infty$. Hence, $\sum_{t=1}^{T} \eta_t a_t$ for $T = 1, 2, ...$ is a martingale in L^2, so it converges in L^2 with probability 1. Overall, with probability 1, the assumptions of Lemma A.1.1 are verified with $p = 2$.

Finally, we proved that the gradients of SGD do indeed converge to zero with probability 1. This means that with probability 1 for any $\epsilon > 0$ there exists N_ϵ such that $\nabla f(w_t) \leq \epsilon$ for $t \geq N_\epsilon$.

Step Size with Cosine Annealing Decay

Before proceeding further, we explore the results for step sizes decaying according to a cosine annealing schedule, $\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\frac{t}{T}\pi\right)\right)$ with $\eta_{min} = 0$. Using cosine annealing step size schedule,

$$\sum_{t=1}^{\infty} \eta_t \to \infty \text{ and } \sum_{t=1}^{\infty} \eta_t^2 < \infty \tag{A.17}$$

Thus, the criteria for convergence still holds for cosine annealing step size schedule. Also, $\sum_{t=1}^{\infty} \eta_t$ grows faster than $\sum_{t=1}^{\infty} \eta_t^2$ However, for finite training periods, $\eta_t = 0$ for $t > T$. Hence, under this condition, we will always have, $\sum_{t=T}^{\infty} \eta_t \mathbb{E}_t \left[\|\nabla f(w_t)\|_2^2\right] = 0 < \infty$ and $\sum_{t=0}^{T} \eta_t \mathbb{E}_t \left[\|\nabla f(w_t)\|_2^2\right] < \infty$. As mentioned before, the aforementioned statement gives rise to a very weak condition for convergence.

In other words, the parameter space $\mathcal{P}$ and the gradient space $\mathcal{G}$, both being a Hausdorff space with complete normed metric $\|\cdot\|$, the sequence of parameters $(\mathcal{P})_{t=1}^{\infty}$ converge within a Ball of radius $r \in \mathbb{R}^{\mathbb{D}}$. Hence, under the assumption of global L-Lipschitz continuity, i.e., $\sum_d h_{\theta n} < \infty$ and $\sum_{w \in \mathcal{P}} w < \infty$, we can infer that $\|\nabla f(x)\|_{t=T} \leq \epsilon$ for $\epsilon > 0$.

References

A. Abdi et al. Pre-training of u-net encoder for improved keypoint detection in transmitral doppler imaging. In *Medical Imaging with Deep Learning (MIDL)*, 2024. URL `https://openreview.net/forum?id=fj5lCxYpCs`.

J. Achiam et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

A. Agrawal, M. Hedlund, and B. Hechtman. exmy: A data type and technique for arbitrary bit precision quantization, 2024.

P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 37–45, 2015.

U. Ahsan, R. Madhok, and I. Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189, 2019.

H. Akbari et al. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 24206–24221, 2021.

M. Akçakaya, B. Yaman, H. Chung, and J. C. Ye. Unsupervised deep learning methods for biological image reconstruction and enhancement: An overview from a signal processing perspective. *IEEE Signal Processing Magazine*, 39(2):28–44, 2022.

H. Alasmawi, L. Bricker, and M. Yaqub. FUSC: fetal ultrasound semantic clustering of second-trimester scans using deep self-supervised learning. *Ultrasound in Medicine and Biology*, 50(5): 703–711, 2024.

Y. I. Alber, A. N. Iusem, and M. V. Solodov. On the projected subgradient method for nonsmooth convex optimization in a hilbert space. *Mathematical Programming*, 81(1):23–35, 1998.

M. Amac, A. Sencan, O. Baran, N. Ikizler-Cinbis, and R. Cinbis. MaskSplit: self-supervised meta-learning for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 428–438, 2022.

N. Araslanov and S. Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15379–15389, 2021.

Y. M. Asano, M. Patrick, C. Rupprecht, and A. Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 4660–4671, 2020a.

Y. M. Asano, C. Rupprecht, and A. Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *8th International Conference on Learning Representations (ICLR)*, 2020b. URL `https://openreview.net/forum?id=Hyx-jyBFPr`.

M. Assran et al. Masked siamese networks for label-efficient learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2022.

M. Assran et al. The hidden uniform cluster prior in self-supervised learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=04K3PMtMckp`.

M. Azabou et al. Mine your own view: Self-supervised learning through across-sample prediction. *CoRR*, abs/2102.10106, 2021. URL `https://arxiv.org/abs/2102.10106`.

S. Azizi et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3458–3468, 2021.

P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 15509–15519, 2019.

H. Bao, L. Dong, and F. Wei. BEiT: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021.

A. Bardes, J. Ponce, and Y. LeCun. VICReg: variance-invariance-covariance regularization for self-supervised learning. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022a. URL `https://openreview.net/forum?id=xm6YD62D1Ub`.

A. Bardes, J. Ponce, and Y. LeCun. VICRegL: self-supervised learning of local visual features. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:8799–8810, 2022b.

D. Basaj et al. Explaining self-supervised image representations with visual probing. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 592–598, 2021.

S. Basu et al. Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 423–433, 2022.

H. Bay, T. Tuytelaars, and L. Van Gool. SURF: speeded up robust features. In *Proceedings of the 8th European Conference on Computer Vision (ECCV) 2006*, pages 404–417, 2006.

S. Bengio and Y. Bengio. Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks*, 11(3):550–557, 2000.

G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824, 2021.

A. K. Bhunia et al. A deep one-shot network for query-based logo retrieval. *Pattern Recognition*, 96:106965, 2019.

N. Bien et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine*, 15(11):e1002699, 2018.

P. Bojanowski and A. Joulin. Unsupervised learning by predicting noise. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 517–526, 2017.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59(4-5):291–294, 1988.

J. Bridle, A. Heading, and D. MacKay. Unsupervised classifiers, mutual information and phantom targets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 4, page 1096–1101, 1991.

T. Brown et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.

H. Buckchash and B. Raman. Sustained self-supervised pretraining for temporal order verification. In *Proceedings of the 8th International Conference on Pattern Recognition and Machine Intelligence (PReMI), Part I*, pages 140–149, 2019.

Z. Cai, L. Lin, H. He, and X. Tang. Uni4Eye: unified 2D and 3D self-supervised pre-training via masked image modeling transformer for ophthalmic image classification. In *Proceedings of the Internation Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 88–98, 2022.

M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the 15th European conference on computer vision (ECCV)*, pages 132–149, 2018.

M. Caron, P. Bojanowski, J. Mairal, and A. Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2959–2968, 2019.

M. Caron et al. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9912–9924, 2020.

M. Caron et al. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

R. B. Cattell. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1, 1963.

K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:12546 – 12558, 2020.

S. Chakraborty, A. Gosthipaty, and S. Paul. G-SimCLR: self-supervised contrastive learning with guided projection via pseudo labelling. In *20th International Conference on Data Mining Workshops, ICDM Workshops*, pages 912–916, 2020.

F. Charte, A. J. Rivera, M. D. Jesús, and F. Herrera. Dealing with difficult minority labels in imbalanced mutilabel data sets. *Neurocomputing*, 326-327:39–53, 2019.

C. Chen, X. Yang, Y. Huang, W. Shi, Y. Cao, M. Luo, X. Hu, L. Zhu, L. Yu, K. Yue, Y. Zhang, Y. Xiong, D. Ni, and W. Huang. FetusMapV2: Enhanced fetal pose estimation in 3D ultrasound. *Medical Image Analysis*, 91(103013):103013, 2024.

J. Chen, X. Zheng, H. Yu, D. Z. Chen, and J. Wu. Electrocardio panorama: Synthesizing new ECG views with self-supervision. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3597–3605, 2021a.

K. Chen, Q. Wang, and Y. Ma. Cervical optical coherence tomography image classification based on contrastive self-supervised texture learning. *Medical Physics*, 49(6):3638–3653, June 2022.

P. Chen, S. Liu, and J. Jia. Jigsaw clustering for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11526–11535, 2021b.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597 – 1607, 2020a.

T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 22243 – 22255, 2020b.

X. Chen and K. He. Exploring simple siamese representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2020.

X. Chen, H. Fan, R. B. Girshick, and K. He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020c. URL `https://arxiv.org/abs/2003.04297`.

X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629, 2021c.

X. Chen, L. Yao, T. Zhou, J. Dong, and Y. Zhang. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern Recognition*, 113:107826, 2021d.

Y. Chen, Z. Yun, Y. Ma, B. A. Olshausen, and Y. LeCun. Minimalistic unsupervised representation learning with the sparse manifold transform. In *The Eleventh International Conference on Learning Representations, (ICLR)*, 2023. URL `https://openreview.net/forum?id=nN_nBVKAhhD`.

J. Chen et al. TransMorph: transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022a.

J. Chen et al. APANet: adaptive prototypes alignment network for few-shot semantic segmentation. *IEEE Transactions on Multimedia*, 25:4361–4373, 2023a.

L. Chen et al. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019. ISSN 1361-8415.

M. Chen et al. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703, 2020.

Q. Chen et al. Colo-SCRL: Self-supervised contrastive representation learning for colonoscopic video retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1056–1061, 2023b.

X. Chen et al. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132:208–223, 2023c.

Z. Chen et al. Masked image modeling advances 3d medical image analysis. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1969–1979, 2022b.

H. Cheng et al. Unsupervised visual representation learning via multi-dimensional relationship alignment. *IEEE Transactions on Image Processing*, 32:1613–1626, 2023.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 539–546, 2005.

P. F. Christiano et al. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 8765–8775, 2020.

O. Ciga, T. Xu, and A. L. Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

F. Cisternino et al. Self-supervised learning for characterising histomorphological diversity and spatial RNA expression prediction across 23 human tissue types. *bioRxiv*, 2023. doi: 10.1101/2023.08.22.554251.

A. Coates and A. Y. Ng. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 561–580. Springer, 2012.

E. Cole, X. S. Yang, K. Wilber, O. M. Aodha, and S. J. Belongie. When does contrastive visual representation learning work? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 01–10, 2021.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pages 2292–2300, 2013.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893, 2005.

P. Damasceno et al. A transthoracic echocardiography (TTE) based PH detection model using view agnostic classifier. *Journal Heart and Lung Transplantation*, 43(4):S406, 2024.

I. R. Dave, R. Gupta, M. N. Rizve, and M. Shah. TCLR: temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2021.

V. De Sa. Learning classification with unlabeled data. *Advances in neural information processing systems (NeurIPS)*, 6, 1993.

V. R. de Sa. Minimizing disagreement for self-supervised classification. In *Proceedings of the 1993 Connectionist Models Summer School*, pages 300–307. Psychology Press, 2014.

O. B. Demirel et al. Improved simultaneous multi-slice functional mri using self-supervised deep learning. In *Proceedings of the 55th Asilomar Conference on Signals, Systems, and Computers*, pages 890–894, 2021.

L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29:141–142, 2012.

J. Deng et al. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, pages 4171–4186, 2019.

T. Devries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. URL `http://arxiv.org/abs/1708.04552`.

F. T. Dezaki et al. Echo-SyncNet: self-supervised cardiac view synchronization in echocardiography. *IEEE Transactions on Medical Imaging*, 40:2092–2104, 2021.

H. Ding, C. Sun, H. Tang, D. Cai, and Y. Yan. Few-shot medical image segmentation with cycle-resemblance attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2487–2496, 2023.

J. Ding et al. An image registration-based self-supervised Su-Net for carotid plaque ultrasound image segmentation. *Computer Methods and Programs in Biomedicine*, 244:107957, 2024.

T.-N. Do. Multi-class bagged proximal support vector machines for the imagenet challenging problem. In *Proceedings of 8th International Conference on Future Data and Security Engineering (FDSE)*, volume 8, pages 99–112. Springer, 2021.

T.-N. Do and H. A. Le Thi. Training support vector machines for dealing with the imagenet challenging problem. In *Proceedings of the 4th International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences (MCO)*, volume 4, pages 235–246. Springer, 2022.

C. Doersch, A. Gupta, and A. A. Efros. Context as supervisory signal: Discovering objects with predictable context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV), Part III 13*, pages 362–377, 2014.

C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

N. Dong and E. P. Xing. Few-shot semantic segmentation with prototype learning. In *British Machine Vision Conference (BMVC)*, page 79, 2018.

N. Dong, M. Kampffmeyer, and I. Voiculescu. Self-supervised multi-task representation learning for sequential medical images. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 779–794, 2021.

X. Dong et al. Bootstrapped masked autoencoders for vision bert pretraining. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, pages 247–264, 2022.

A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 766–774, 2014.

A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

H. Duan, N. Zhao, K. Chen, and D. Lin. Transrank: Self-supervised video representation learning via ranking-based transformation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3000–3010, 2022.

D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9568–9577, 2021.

A. El-Nouby, S. Zhai, G. W. Taylor, and J. M. Susskind. Skip-clip: Self-supervised spatiotemporal representation learning by future clip order ranking. *arXiv preprint arXiv:1910.12770*, 2019.

A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe. Whitening for self-supervised representation learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 3015–3024, 2021.

Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang. Self-support few-shot semantic segmentation. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, pages 701–719, 2022.

Z. Fan et al. Medical image classification using self-supervised learning-based masked autoencoder. In *Medical Imaging 2024: Image Processing*, volume 12926, pages 123–129, 2024.

Y. Fang, L. Dong, H. Bao, X. Wang, and F. Wei. Corrupted image modeling for self-supervised visual pre-training. In *The Eleventh International Conference on Learning Representations, (ICLR)*, 2023. URL `https://openreview.net/pdf?id=09hVcSDkea`.

Y. Fang et al. A self-supervised classification model for endometrial diseases. *Journal of Cancer Research and Clinical Oncology*, 149(20):17855–17863, 2023.

Z. Fei et al. Masked auto-encoders meet generative adversarial networks and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24449–24459, 2023.

C. Feichtenhofer, H. Fan, Y. Li, and K. He. Masked autoencoders as spatiotemporal learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 35946–35958, 2022.

P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, Sept. 2004.

Z. Feng, C. Xu, and D. Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10364–10374, 2019.

B. Fernando, H. Bilen, E. Gavves, and S. Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3636–3645, 2017.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2): 179–188, 1936.

K. Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

A. Galdran, J. W. Verjans, G. Carneiro, and M. A. González Ballester. Multi-head multi-loss model calibration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 108–117, 2023.

P. Gallinari, Y. Lecun, S. Thiria, and F. Fogelman Soulie. Memoires associatives distribuees: Une comparaison (distributed associative memories: A comparison). In *Proceedings of COGNITIVA 87*, 1987.

T. Gan et al. Self-supervised representation learning using feature pyramid siamese networks for colorectal polyp detection. *Scientific Reports*, 13(1):21655, 2023.

W. V. Gansbeke, S. Vandenhende, S. Georgoulis, and L. V. Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10032–10042, 2021.

P. Gao et al. MCMAE: masked convolution meets masked autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 35632 – 35644, 2022a.

Z. Gao et al. Unsupervised representation learning for tissue segmentation in histopathological images: From global to local contrast. *IEEE Transactions on Medical Imaging*, 41(12):3611–3623, 2022b.

C. Ge et al. Soft neighbors are positive supporters in contrastive visual representation learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/pdf?id=l9vM_PaUKz`.

S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, (ICLR) 2018*, 2018.

S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, volume 3056 of *Lecture Notes in Computer Science*, pages 22–30, 2004.

M. Goncharov, V. Soboleva, A. Kurmukov, M. Pisov, and M. Belyaev. vox2vec: A framework for self-supervised contrastive learning of voxel-level representations in medical images. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 605–614, 2023.

I. Goodfellow et al. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

P. Goyal et al. Vision models are more robust and fair when pretrained on uncurated images without supervision. *ArXiv*, abs/2202.08360, 2022.

D. Granziol, X. Wan, and T. Garipov. Deep curvature suite, 2019.

A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf. Measuring statistical dependence with hilbert-schmidt norms. In *16th International Conference on Algorithmic Learning Theory*, pages 63–77, 2005.

J.-B. Grill et al. Bootstrap your own latent a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21271 – 21284, 2020.

A. Guha Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger. 'Squeeze & excite' guided few-shot segmentation of volumetric images. *Medical Image Analysis*, 59:101587, 2020.

V. Guizilini and F. Ramos. Online self-supervised segmentation of dynamic objects. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4720–4727, 2013.

X. Guo, L. Gao, X. Liu, and J. Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1753–1759, 2017.

E. Gupta, V. Gupta, M. Chopra, P. C. Chhipa, and M. Liwicki. Learning self-supervised representations for label efficient cross-domain knowledge transfer on diabetic retinopathy fundus images. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2023.

D. Gutman et al. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *CoRR*, abs/1605.01397, 2016. URL `http://arxiv.org/abs/1605.01397`.

M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research (JMLR)*, 13(11):307–361, 2012.

M. Gwilliam and A. Shrivastava. Beyond supervised vs. unsupervised: Representative benchmarking and analysis of image representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9632–9642, 2022.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742, 2006.

F. Haghighi, M. R. Hosseinzadeh Taher, Z. Zhou, M. B. Gotway, and J. Liang. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 137–147, 2020.

F. Haghighi, M. Taher, M. B. Gotway, and J. Liang. DiRA: discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20792–20802, 2022.

F. Haghighi, M. R. H. Taher, M. B. Gotway, and J. Liang. Self-supervised learning for medical image analysis: Discriminative, restorative, or adversarial? *Medical Image Analysis*, 94:103086, 2024.

M. S. Halvagal, A. Laborieux, and F. Zenke. Implicit variance regularization in non-contrastive SSL. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

T. Han, W. Xie, and A. Zisserman. Self-supervised co-training for video representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 5679 – 5690, 2020.

J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 5000–5011, 2021.

B. He and M. Ozay. Exploring the gap between collapsed and whitened features in self-supervised learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 8613–8634, 2022.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

H. He, J. Zhang, B. Thuraisingham, and D. Tao. Progressive one-shot human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1522–1530, 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

K. He et al. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022.

O. J. Hénaff et al. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, ICML'20, 2020.

Á. S. Hervella, J. Rouco, J. Novo, and M. Ortega. Retinal image understanding emerges from self-supervised multimodal reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 321–328, 2018.

Á. S. Hervella, L. Ramos, J. Rouco, J. Novo, and M. Ortega. Multi-modal self-supervised pre-training for joint optic disc and cup segmentation in eye fundus images. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 961–965, 2020.

G. E. Hinton and R. S. Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 6, page 3–10, 1993.

G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL `http://arxiv.org/abs/1503.02531`.

R. D. Hjelm et al. Learning deep representations by mutual information estimation and maximization. In *The 7th International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=Bklr3j0cKX`.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840 – 6851, 2020.

O. G. Holmberg et al. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. *Nature Machine Intelligence*, 2 (11):719–726, 2020.

L. T. T. Hong, N. C. Thanh, and T. Q. Long. Self-supervised visual feature learning for polyp segmentation in colonoscopy images using image reconstruction as pretext task. In *8th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 254–259, 2021.

J. L. Horn. *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities*. University of Illinois at Urbana-Champaign, 1965.

J. L. Horn and R. B. Cattell. Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of educational psychology*, 57(5):253, 1966.

L. Hoyer et al. Three ways to improve semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11125–11135, 2021.

S. Hu, C. Zhang, G. Zou, Z. Lou, and Y. Ye. Deep multiview clustering by pseudo-label guided contrastive learning and dual correlation learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2024.

W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Research*, page 1558–1567, 2017.

S.-Y. Hu et al. Self-supervised pretraining with DICOM metadata in ultrasound imaging. In *Proceedings of the 5th Machine Learning for Healthcare Conference (MLHC)*, volume 126 of *Proceedings of Machine Learning Research*, pages 732–749, 2020.

T. Hua et al. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9578–9588, 2021.

J. Huang, H. Li, G. Li, and X. Wan. Attentive symmetric autoencoder for brain MRI segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 203–213, 2022.

Z. Huang, J. Chen, J. Zhang, and H. Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:7509–7524, 2021.

D. Huang et al. ASCNet: self-supervised video representation learning with appearance-speed consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8076–8085, 2021.

Z. Huang et al. Model-aware contrastive learning: Towards escaping the dilemmas. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 13774–13790, 2023a.

Z. Huang et al. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2506–2517, 2023b.

Y. Huo et al. Compressed video contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 14176–14187, 2021.

T. Huynh, A. Nibali, and Z. He. Semi-supervised learning for medical image classification using imbalanced training data. *Computer Methods and Programs in Biomedicine*, 216:106628, 2022.

A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 3765–3773, 2016.

S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transaction on Graphs*, 35(4):1–11, jul 2016.

Y. Intrator, N. Aizenberg, A. Livne, E. Rivlin, and R. Goldenberg. Self-supervised polyp re-identification in colonoscopy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Part V*, volume 14224 of *Lecture Notes in Computer Science*, pages 590–600, 2023.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pages 448–456, 2015.

ISDIS. https://challenge.isic-archive.com/leaderboards/2016/, 2016. URL `https://challenge.isic-archive.com/leaderboards/2016/`.

A. Islam et al. A broad study on the transferability of visual representations with contrastive learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8825–8835, 2021.

A. Jamaludin, T. Kadir, and A. Zisserman. Self-supervised learning for spinal MRIs. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 294–302, 2017.

A. Jana et al. Liver fibrosis and nas scoring from ct images using self-supervised learning and texture encoding. In *Proceedings of the IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1553–1557, 2021.

D. Jayaraman and K. Grauman. Learning image representations tied to egomotion from unlabeled video. *International Journal of Computer Vision*, 125:136–161, 2017.

S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

X. Ji, A. Vedaldi, and J. Henriques. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9864–9873, 2019.

G.-Z. Jian, G.-S. Lin, C.-M. Wang, and S.-L. Yan. Helicobacter pylori infection classification based on convolutional neural network and self-supervised learning. In *Proceedings of the 5th International Conference on Graphics and Signal Processing (ICGSP)*, page 60–64, 2021.

Y. Jiang et al. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15813–15823, 2023.

J. Jiao, Y. Cai, M. Alsharid, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Self-supervised contrastive video-speech representation learning for ultrasound. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 12263:534–543, 2020.

J. Jiao et al. Self-supervised representation learning for ultrasound video. *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2020:1847–1850, 2020.

J. Jiao et al. Show from tell: Audio-visual modelling in clinical settings. *CoRR*, abs/2310.16477, 2023.

L. Jing and Y. Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *CoRR*, abs/1811.11387, 2018.

A. Jog, A. Carass, and J. L. Prince. Self super-resolution for magnetic resonance images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Part III*, pages 553–560, 2016.

Q. Kang, J. Gao, K. Li, and Q. Lao. Deblurring masked autoencoder is better recipe for ultrasound image recognition. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 352–362, 2023.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Part I 16*, pages 795–811, 2016.

A. E. Kavur et al. CHAOS Challenge - combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

P. Khosla et al. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18661–18673, 2020.

D. Kim, D. Cho, D. Yoo, and I. S. Kweon. Learning image representations by completing damaged jigsaw puzzles. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 793–802, 2018.

D. Kim, D. Cho, and I. S. Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 8545–8552, 2019.

S. Kim, G. Lee, S. Bae, and S.-Y. Yun. MixCo: mix-up contrastive learning for visual representation. *arXiv preprint arXiv:2010.06300*, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL `http://arxiv.org/abs/1412.6980`.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014. URL `http://arxiv.org/abs/1312.6114`.

A. Kolesnikov, X. Zhai, and L. Beyer. Revisiting self-supervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1920–1929, 2019.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN 0262013193.

S. A. Koohpayegani, A. Tejankar, and H. Pirsiavash. Mean shift for self-supervised learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10306–10315, 2021.

B. Korbar, D. Tran, and L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems (NeurIPS)*, 31: 7774–7785, 2018.

N. Kouiroukidis and G. Evangelidis. The effects of dimensionality curse in high dimensional knn search. In *Proceedings of the 15th Panhellenic Conference on Informatics*, pages 41–45, 2011.

A. Krause, P. Perona, and R. Gomes. Discriminative clustering by regularized information maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23, page 775–783, 2010.

A. Krizhevsky. Learning multiple layers of features from tiny images, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NeurIPS)*, volume 25, 2012.

A. Kukleva, M. Böhle, B. Schiele, H. Kuehne, and C. Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *The Eleventh International Conference on Learning Representations, (ICLR)*, 2023. URL `https://openreview.net/forum?id=ejHUr4nfHhD`.

V. Kumar, V. Tripathi, and B. Pant. Unsupervised learning of visual representations via rotation and future frame prediction for video retrieval. In *Advances in Computing and Data Sciences (ICACDS), Part I*, volume 5, pages 701–710, 2021.

A. Kuzmin et al. FP8 quantization: The power of the exponent. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14651–14662, 2022.

E. Lamoureux, S. Ayromlou, S. N. Ahmadi Amiri, and H. Rhodin. Segmenting cardiac ultrasound videos using self-supervised learning. In *45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7, 2023.

G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. In *Proceedings of the European Conference on Computer Vision (ECCV), Part IV 14*, pages 577–593, 2016.

G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6874–6883, 2017.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Y. LeCun et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989a.

Y. LeCun et al. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 2, pages 396–404, 1989b.

D. H. Lee, S. Choi, H. J. Kim, and S. Chung. Unsupervised visual representation learning via mutual information regularized assignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 29610–29623, 2022.

H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 667–676, 2017.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, 2016.

J. Lee et al. Self supervised convolutional kernel based handcrafted feature harmonization: Enhanced left ventricle hypertension disease phenotyping on echocardiography. *ArXiv*, abs/2310.08897, 2023.

K. Lee et al. i-Mix: A domain-agnostic strategy for contrastive representation learning. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=T6AxtOaWydQ`.

B. Li, Y. Li, and K. W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, 2020.

J. Li, P. Zhou, C. Xiong, and S. C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *9th International Conference on Learning Representations (ICLR)*, 2021a. URL `https://openreview.net/forum?id=KmykpuSrjcq`.

R. Li, S. Liu, G. Wang, G. Liu, and B. Zeng. JigsawGAN: auxiliary learning for solving jigsaw puzzles with generative adversarial networks. *IEEE Transactions on Image Processing*, 31: 513–524, 2022.

X. Li, X. Hu, X. Qi, L. Yu, W. Zhao, P.-A. Heng, and L. Xing. Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 40:2284–2294, 2021b.

Y. Li, R. Pogodin, D. J. Sutherland, and A. Gretton. Self-supervised learning with kernel dependence maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15543 – 15556, 2021c.

Y. Li, Y. Huang, N. He, K. Ma, and Y. Zheng. Improving vision transformer for medical image classification via token-wise perturbation. *Journal of Visual Communication and Image Representation*, 98:104022, 2024.

C. Li et al. Efficient self-supervised vision transformers for representation learning. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=fVu3o-YUGQK`.

G. Li et al. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8330–8339, 2021a.

H. Li et al. Imbalance-aware self-supervised learning for 3d radiomic representations. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 36–46, 2021b.

J. Li et al. DSMT-Net: Dual self-supervised multi-operator transformation for multi-source endoscopic ultrasound diagnosis. *IEEE Transactions on Medical Imaging*, 43(1):64–75, 2024.

Y. Li et al. Cross-shaped windows transformer with self-supervised pretraining for clinically significant prostate cancer detection in bi-parametric MRI. *CoRR*, abs/2305.00385, 2023. URL `https://doi.org/10.48550/arXiv.2305.00385`.

Z. Li et al. MST: masked self-supervised transformer for visual representation. In *Advances in Neural Information Processing Systems*, volume 34, pages 13165–13176, 2021c.

H. Liang, G. Ning, X. Zhang, and H. Liao. Semi-supervised anatomy tracking with contrastive representation learning in ultrasound sequences. In *Proceedings of the IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023.

H. Liang et al. Self-supervised spatiotemporal representation learning by exploiting video continuity. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pages 1564–1573, 2022.

S. Liang et al. Skin lesion classification base on multi-hierarchy contrastive learning with pareto optimality. *Biomedical Signal Processing and Control*, 86:105187, 2023.

Z. Lin, R. Huang, D. Ni, J. Wu, and B. Luo. Masked video modeling with correlation-aware contrastive learning for breast cancer diagnosis in ultrasound. In *MICCAI Workshop on Resource-Efficient Medical Image Analysis*, pages 105–114, 2022.

T.-Y. Lin et al. Microsoft COCO: Common objects in context. *CoRR*, abs/1405.0312, 2014.

Y. Liu, X. Zhang, S. Zhang, and X. He. Part-aware prototype network for few-shot semantic segmentation. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pages 142–158, 2020.

H. Liu et al. M3AE: multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1657–1665, 2023.

J. Liu et al. Dynamic prototype convolution network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11543–11552, 2022a.

W. Liu et al. Few-shot segmentation with optimal transport matching and message flow. *IEEE Transactions on Multimedia*, 25:5130–5141, 2022b.

Z. Liu et al. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022c.

Z. Liu et al. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2022d.

D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, 1999.

Q. Lu, Y. Li, and C. Ye. White matter tract segmentation with self-supervised learning. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 270–279, 2020.

Q. Lu, Y. Li, and C. Ye. Volumetric white matter tract segmentation with nested self-supervised learning using sequential pretext tasks. *Medical Image Analysis*, 72:102094, 2021.

X. Lu et al. Self-supervised dual-head attentional bootstrap learning network for prostate cancer screening in transrectal ultrasound images. *Computers in Biology and Medicine*, 165:107337, 2023.

D. Luo and J. Wang. Prior matching operator in self-supervised learning. In *Proceedings of the 7th International Conference on Signal and Image Processing (ICSIP)*, pages 777–781, 2022. doi: 10.1109/ICSIP55141.2022.9886345.

D. Luo et al. Video cloze procedure for self-supervised spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11701–11708, 2020.

S. Ma, Z. Zeng, D. McDuff, and Y. Song. Active contrastive learning of audio-visual video representations. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=OMizHuea_HB`.

F. Maani, A. Ukaye, N. Saadi, N. Saeed, and M. Yaqub. SimLVSeg: simplifying left ventricular segmentation in 2D+Time echocardiograms with self- and weakly-supervised learning, 2024. URL `https://arxiv.org/abs/2310.00454`.

D. Mahapatra, A. Poellinger, L. Shao, and M. Reyes. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2548–2562, 2021.

J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, page 2283–2291, 2013.

S. Manna, S. Bhattacharya, and U. Pal. Interpretive self-supervised pre-training: boosting performance on visual medical data. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021a.

S. Manna, S. Bhattacharya, and U. Pal. MIO : Mutual information optimization using self-supervised binary contrastive learning. *CoRR*, abs/2111.12664, 2021b. URL `https://arxiv.org/abs/2111.12664`.

S. Manna, S. Bhattacharya, and U. Pal. Self-supervised representation learning for detection of ACL tear injury in knee MR videos. *Pattern Recognition Letters*, 154:37–43, 2022.

S. Manna, S. Bhattacharya, and U. Pal. Self-supervised representation learning for knee injury diagnosis from magnetic resonance data. *IEEE Transactions on Artificial Intelligence*, 5(4): 1613–1623, 2023.

S. Manna, S. Chattopadhyay, R. Dey, S. Bhattacharya, and U. Pal. Dynamically scaled temperature in self-supervised contrastive learning. *arXiv preprint arXiv:2308.01140*, 2024.

J. S. Marron, M. J. Todd, and J. Ahn. Distance-Weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.

M. Mazher et al. Self-supervised spatial–temporal transformer fusion based federated framework for 4d cardiovascular image segmentation. *Information Fusion*, 106:102256, 2024.

A. Maćkiewicz and W. Ratajczak. Principal components analysis (pca). *Computers and Geosciences*, 19(3):303–342, 1993.

D. McAllester and K. Stratos. Formal limitations on the measurement of mutual information. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 875–884, 2020.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

L. McInnes, J. Healy, and J. Melville. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426, 2018. URL `http://arxiv.org/abs/1802.03426`.

T. Mehari and N. Strodthoff. Self-supervised representation learning from 12-lead ECG data. *Computers in Biology and Medicine*, 141(C), 2022.

P. Micikevicius et al. Fp8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022.

A. Miech et al. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, 2020.

I. Misra and L. van der Maaten. Self-supervised learning of pretext-invariant representations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716, 2019.

I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Proceedings of the 14th European Conference on Computer Vision (ECCV), Part I 14*, pages 527–544. Springer, 2016.

J. Mitrovic, B. McWilliams, J. C. Walker, L. H. Buesing, and C. Blundell. Representation learning via invariant causal mechanisms. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=9p2ekP904Rs`.

N. Mojab et al. Real-world multi-domain data applications for generalizations to clinical settings. *Proceedings of the 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 677–684, 2020.

A. K. Monsefi et al. Masked LoGoNet: Fast and accurate 3d image analysis for medical domain, 2024.

J. H. Moon, W. Kim, and E. Choi. Correlation between alignment-uniformity and performance of dense contrastive representations. In *33rd British Machine Vision Conference (BMVC)*, page 844, 2022.

P. Morgado, N. Vasconcelos, and I. Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12470–12481, 2021.

T. Morita and X.-H. Han. Investigating self-supervised learning for skin lesion classification. In *Proceedings of the 18th International Conference on Machine Vision and Applications (MVA)*, pages 1–5, 2023.

M. Nauta, J. H. Hegeman, J. Geerdink, J. Schlötterer, M. v. Keulen, and C. Seifert. Interpreting and correcting medical image classification with PIP-Net. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 198–215, 2023a.

M. Nauta, J. Schlötterer, M. van Keulen, and C. Seifert. PIP-Net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2744–2753, 2023b.

K. L. Navaneet et al. Constrained mean shift using distant yet related neighbors for representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV), Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 23–41, 2022.

N.-Q. Nguyen and T.-S. Le. A semi-supervised learning method to remedy the lack of labeled data. In *15th International Conference on Advanced Computing and Applications (ACOMP)*, pages 78–84, 2021.

D. M. Nguyen et al. Joint self-supervised image-volume representation learning with intra-inter contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14426–14435, 2023.

D. T. Nguyen et al. DeepUSPS: deep robust unsupervised saliency prediction via self-supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 204–214, 2019.

M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1447–1454, 2006.

M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Sixth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pages 722–729, 2008.

M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pages 69–84. Springer, 2016.

M. Noroozi, H. Pirsiavash, and P. Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5899–5907, 2017.

A. Okazawa. Interclass prototype relation for few-shot segmentation. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, pages 362–378, 2022.

M. Oquab et al. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023.

F. Orabona. Almost sure convergence of sgd on smooth non-convex functions. `https://parameterfree.com/2020/10/05/almost-sure-convergence-of-sgd-on-smooth-non-convex-functions/`, 2020. [Accessed 25-04-2024].

E. Orhan, V. Gupta, and B. M. Lake. Self-supervised learning through the eyes of a child. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9960–9971, 2020.

Y. Ouali, C. Hudelot, and M. Tami. Autoregressive unsupervised image segmentation. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pages 142–158, 2020.

C. Ouyang et al. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pages 762–780, 2020.

C. Ouyang et al. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022.

J. Ouyang et al. Self-supervised longitudinal neighbourhood embedding. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 80–89, 2021.

T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu. VideoMoCo: contrastive video representation learning with temporally adversarial examples. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11200–11209, 2021.

B. Pang, Y. Zhang, Y. Li, J. Cai, and C. Lu. Unsupervised visual representation learning by synchronous momentum grouping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–282, 2022.

R. Pascanu, Y. N. Dauphin, S. Ganguli, and Y. Bengio. On the saddle point problem for non-convex optimization. *CoRR*, abs/1405.4604, 2014. URL `http://arxiv.org/abs/1405.4604`.

V. Patel and A. S. Berahas. Gradient descent in the absence of global lipschitz continuity of the gradients. *SIAM Journal on Mathematics of Data Science*, 6(3):602–626, 2024.

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2536–2544, 2016.

M. Patrick et al. Space-time crop and attend: Improving cross-modal video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10540–10552, 2021.

S. Perek, M. Amit, and E. Hexter. Self supervised contrastive learning on multiple breast modalities boosts classification performance. In *4th International Workshop on Predictive Intelligence in Medicine (PRIME), Held in Conjunction with MICCAI*, volume 12928, pages 117–127, 2021.

M. Pihlaja, M. Gutmann, and A. Hyvärinen. A family of computationally efficient and simple estimators for unnormalized statistical models. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, page 442–449, 2010.

M. Prakash et al. Leveraging self-supervised denoising for image segmentation. In *Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 428–432, 2020.

N. S. Punn and S. Agarwal. BT-Unet: a self-supervised learning framework for biomedical image segmentation using barlow twins with u-net models. *Machine Learning*, 111(12):4585–4600, 2022.

H. Qi, S. Collins, and J. A. Noble. Knowledge-guided pretext learning for utero-placental interface detection. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 582–593, 2020.

Q. Qian, Y. Xu, J. Hu, H. Li, and R. Jin. Unsupervised visual representation learning by online constrained k-means. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16619–16628, 2021.

R. Qian et al. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6960–6970, 2021.

Z. Qing et al. MAR: masked autoencoders for efficient action recognition. *IEEE Transactions on Multimedia*, 26:218–233, 2023.

Z. Qiu et al. Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. In *Proceedings of the International Conference on Machine Learning (ICML) 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 28389–28421, 2023.

A. Radford. Improving language understanding by generative pre-training, 2018.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations (ICLR)*, 2016. URL `http://arxiv.org/abs/1511.06434`.

A. Radford et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

P. Rajpurkar et al. MURA dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. *CoRR*, abs/1712.06957, 2017. URL `http://arxiv.org/abs/1712.06957`.

K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine. Conditional networks for few-shot semantic segmentation. In *Workshop Track Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018a. URL `https://openreview.net/forum?id=SkMjFKJwG`.

K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine. Few-shot segmentation propagation with guided networks. *ArXiv*, abs/1806.07373, 2018b.

A. Ramesh et al. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831, 2021.

A. Recasens et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1235–1245, 2021.

P. H. Richemond et al. Byol works even without batch statistics. *ArXiv*, abs/2010.10241, 2020.

J. D. Robinson, C. Chuang, S. Sra, and S. Jegelka. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=CR1XOQ0UTh-`.

J. T. Rolfe. Discrete variational autoencoders. In *5th International Conference on Learning Representations (ICLR)*, 2017. URL `https://openreview.net/forum?id=ryMxXPFex`.

B. W. Roop, K. J. Brady, L. A. Gjesteby, B. S. Baum, and L. J. Brattain. Self-supervised learning for ultrasound probe angle prediction in plantar fascia images. In *Proceedings of the IEEE 19th International Conference on Body Sensor Networks (BSN)*, pages 1–6, 2023.

T. Ross et al. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery*, 13:925–933, 2017.

C. Ryali, D. J. Schwab, and A. S. Morcos. Learning background invariance improves generalization and robustness in self-supervised learning on imagenet and beyond. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021. URL `https://openreview.net/forum?id=zZnOG9ehfoO`.

A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou. Spreading vectors for similarity search. In *7th International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=SkGuG2R5tm`.

C. Saillard et al. Self supervised learning improves dMMR/MSI detection from histology slides across multiple cancers. In *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 156 of *Proceedings of Machine Learning Research*, pages 191–205, 2021.

Q. Sang, Y. Hou, P. Qian, and Q. Wu. Self-supervised learning-leveraged boosting ultrasound image segmentation via mask reconstruction. *International Journal of Machine Learning and Cybernetics*, Nov. 2023.

L. K. Saul and S. T. Roweis. An introduction to locally linear embedding, 2001. URL `https://cs.nyu.edu/~roweis/lle/papers/lleintro.pdf`.

R. R. Selvaraju, K. Desai, J. Johnson, and N. Naik. CASTing Your Model: learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11053–11062, 2021.

R. R. Selvaraju et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

P. Sermanet, C. Lynch, J. Hsu, and S. Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 486–487, 2017.

J. A. Serret. *Cours de calcul différentiel et intégral*. Gauthier-Villars, Imprimeur-Libraire, 1868.

A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots. One-shot learning for semantic segmentation. In *British Machine Vision Conference (BMVC)*, 2017. URL `https://www.dropbox.com/s/1odhw88t465klsz/0797.pdf?dl=1`.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, 1948.

C. Shen et al. Asymmetric patch sampling for contrastive learning. *CoRR*, abs/2306.02854, 2023. URL `https://doi.org/10.48550/arXiv.2306.02854`.

Z. Shen et al. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2216–2224, 2022.

X. Shi et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, page 802–810, 2015.

M. Siam and B. N. Oreshkin. Adaptive masked weight imprinting for few-shot segmentation. In *Workshop at the International Conference on Learning Representations (ICLR)*, 2019.

M. Siam, B. Oreshkin, and M. Jagersand. AMP: adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5248–5257, 2019.

F. Siar, A. Gheibi, and A. Mohades. Unsupervised learning of visual representations by solving shuffled long video-frames temporal order prediction. In *ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference (ACM SIGGRAPH), Posters*, pages 1–2, 2020.

S. Singh et al. Self-supervised feature learning for semantic segmentation of overhead imagery. In *British Machine Vision Conference (BMVC)*, page 102, 2018.

K. Sohn. Improved deep metric learning with multi-class N-pair loss objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 1849–1857, 2016.

H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning (MIDL)*, volume 143 of *Proceedings of Machine Learning Research*, pages 728–744, 2021.

H. Spitzer, K. Kiwitz, K. Amunts, S. Harmeling, and T. Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 663–671, 2018.

M. Springenberg et al. From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. *Medical Image Analysis*, 87:102809, 2023.

K. Stacke, C. Lundström, J. Unger, and G. Eilertsen. Evaluation of contrastive predictive coding for histopathology applications. In E. Alsentzer, M. B. A. McDermott, F. Falck, S. K. Sarkar, S. Roy, and S. L. Hyland, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pages 328–340, 2020.

I. Štajduhar, M. Mamula, D. Miletić, and G. Ünal. Semi-automated detection of anterior cruciate ligament injury from MRI. *Computer Methods and Programs in Biomedicine*, 140:151–164, 2017.

T. Stegmüller, T. Lebailly, B. Bozorgtabar, T. Tuytelaars, and J.-P. Thiran. CrOC: cross-view online clustering for dense visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7000–7009, 2023.

D. Štepec and D. Skočaj. Image synthesis as a pretext for unsupervised histopathological diagnosis. In *Proceedings of the 5th International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), Held in Conjunction with MICCAI 2020, Proceedings 5*, pages 174–183, 2020.

C. Sun, F. Baradel, K. Murphy, and C. Schmid. Contrastive bidirectional transformer for temporal representation learning. *CoRR*, abs/1906.05743, 2019. URL `http://arxiv.org/abs/1906.05743`.

L. Sun, K. Yu, and K. Batmanghelich. Context matters: Graph-based self-supervised representation learning for medical images. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 35(6):4874–4882, 2021.

W. Sun et al. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6420–6429, 2023.

Y. Sun et al. Multi-scale self-supervised learning for multi-site pediatric brain mr image segmentation with motion/gibbs artifacts. *Machine learning in medical imaging. MLMI*, 12966:171–179, 2021.

I. Susmelj et al. Lightly. *GitHub. Note: https://github.com/lightly-ai/lightly*, 2020.

C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 4278–4284, 2017.

R. Tadokoro, R. Yamada, and H. Kataoka. Pre-training auto-generated volumetric shapes for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4740–4745, 2023a.

R. Tadokoro, R. Yamada, K. Nakashima, R. Nakamura, and H. Kataoka. Primitive geometry segment pre-training for 3d medical image segmentation. In *34th British Machine Vision Conference 2022, (BMVC)*, pages 152–160, 2023b.

M. R. H. Taher, F. Haghighi, M. B. Gotway, and J. Liang. CAiD: Context-Aware instance discrimination for self-supervised learning in medical imaging. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, pages 535–551, 2022.

A. Taleb, C. Lippert, T. Klein, and M. Nabi. Multimodal self-supervised learning for medical image analysis. In *Information Processing in Medical Imaging*, pages 661–673, 2021.

A. Tejankar, S. A. Koohpayegani, V. Pillai, P. Favaro, and H. Pirsiavash. ISD: self-supervised learning by iterative similarity distillation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9589–9598, 2020.

A. Tejankar, S. A. Koohpayegani, and H. Pirsiavash. Constrained mean shift for representation learning. *CoRR*, abs/2110.10309, 2021. URL `https://arxiv.org/abs/2110.10309`.

Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part XI*, pages 776–794, 2020.

Y. Tian, X. Chen, and S. Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *Proceedings of the 38th International Conference on Machine Learning, (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 10268–10278, 2021.

E. Tiu et al. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

M. N. N. To et al. LensePro: label noise-tolerant prototype-based network for improving cancer detection in prostate ultrasound with limited annotations. *International Journal of Computer Assisted Radiology and Surgery*, 2024.

Z. Tong, Y. Song, J. Wang, and L. Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 10078 – 10093, 2022.

Y. H. Tsai et al. Self-supervised representation learning with relative predictive coding. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=068E_JSq9O`.

G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3:1–13, 2007.

A. Unwin and K. Kleinman. The iris data set: In search of the source of virginica. *Significance*, 18(6):26–29, 2021.

V. Useini et al. Automatized self-supervised learning for skin lesion screening. *Scientific Reports*, 14(1):12697, 2024.

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(86):2579–2605, 2008.

W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool. SCAN: learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–285, 2020.

B. VanBerlo, A. Wong, J. Hoey, and R. Arntfield. Intra-video positive pairs in self-supervised learning for ultrasound. *Frontiers in Imaging*, 3, 2024.

A. Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017.

A. Vats, M. Pedersen, and A. Mohammed. A preliminary analysis of self-supervision for wireless capsule endoscopy. In *9th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6, 2021.

Y. N. T. Vu et al. MedAug: contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In *Proceedings of the 6th Machine Learning for Healthcare Conference (MLHC)*, volume 149 of *Proceedings of Machine Learning Research*, pages 755–769, 2021.

D. Wang, N. Pang, Y. Wang, and H. Zhao. Unlabeled skin lesion classification by self-supervised topology clustering network. *Biomedical Signal Processing and Control*, 66:102428, 2021a.

F. Wang and H. Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021*, pages 2495–2504, 2021.

F. Wang, T. Kong, R. Zhang, H. Liu, and H. Li. Self-supervised learning by estimating twin class distributions. *CoRR*, abs/2110.07402, 2021b. URL `https://arxiv.org/abs/2110.07402`.

H. Wang, E. Ahn, L. Bi, and J. Kim. Self-supervised multi-modality learning for multi-label skin lesion classification. *arXiv preprint arXiv:2310.18583*, 2023.

K. Wang, J. Liew, Y. Zou, D. Zhou, and J. Feng. PANet: few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9196–9205, 2019.

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML) 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939, 2020.

X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2015.

X. Wang and G.-J. Qi. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:5549–5560, 2021.

X. Wang, K. He, and A. Gupta. Transitive invariance for self-supervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1338–1347, 2017.

X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li. Dense contrastive learning for self-supervised visual pre-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3023–3032, 2020.

H. Wang et al. Few-shot semantic segmentation with democratic attention networks. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pages 730–746, 2020.

J. Wang et al. SSL2: self-supervised learning meets semi-supervised learning: multiple sclerosis segmentation in 7t-mri from large-scale 3t-mri. In *Medical Imaging*, 2023a.

J. Wang et al. Thyroid ultrasound diagnosis improvement via multi-view self-supervised learning and two-stage pre-training. *Computers in Biology and Medicine*, 171(108087):108087, 2024.

L. Wang et al. RePre: improving self-supervised vision transformer with reconstructive pre-training. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1437–1443, 2022a.

L. Wang et al. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023b.

R. Wang et al. BEVT: BERT pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14713–14723, 2022b.

Z. Wang et al. Exploring set similarity for dense self-supervised representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16569–16578, 2021.

Z. Wang et al. SSCAP: Self-supervised co-occurrence action parsing for unsupervised temporal action segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 175–184, 2022c.

C. Wei, H. Wang, W. Shen, and A. L. Yuille. CO2: consistent contrast for unsupervised visual representation learning. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=U4XLJhqwNF1`.

C. Wei et al. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1910–1919, 2019.

C. Wei et al. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14648–14658, 2022.

C. Wei et al. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16284–16294, October 2023.

J. W. Wei et al. A petri dish for histopathology image analysis. In *Proceedings of the 19th International Conference on Artificial Intelligence in Medicine (AIME)*, volume 12721, pages 11–24, 2021.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research (JMLR)*, 10:207–244, 2009. ISSN 1532-4435.

R. Windsor, A. Jamaludin, T. Kadir, and A. Zisserman. Self-supervised multi-modal alignment for whole body medical imaging. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 90–101, 2021.

S. Wolf, M. Lalit, H. Westmacott, K. McDole, and J. Funke. Unsupervised learning of object-centric embeddings for cell instance segmentation in microscopy images. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21206–21215, 2023.

S. Woo et al. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142, June 2023.

H. Wu, F. Xiao, and C. Liang. Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, pages 417–434, 2022a.

Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu. Federated contrastive learning for volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 367–377, 2021.

Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu. Distributed contrastive learning for medical image segmentation. *Medical Image Analysis*, 81:102564, 2022b.

Y. Wu, B. Zheng, J. Chen, D. Z. Chen, and J. Wu. Self-learning and one-shot learning based single-slice annotation for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 244–254, 2022c.

Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018.

Q. Wu et al. Denoising masked autoencoders help robust classification. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/pdf?id=zDjtZZBZtqK`.

J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, volume 48, page 478–487, 2016.

Y. Xie et al. Identification method of thyroid nodule ultrasonography based on self-supervised learning dual-branch attention learning framework. *Health Information Science and Systems*, 12 (1):7, 2024.

Z. Xie et al. SimMIM: a simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9643–9653, 2022.

Y. Xiong, M. Ren, W. Zeng, and R. Waabi. Self-supervised representation learning from flow equivariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10171–10180, 2021.

J. Xu, E. Abaci Turk, P. E. Grant, P. Golland, and E. Adalsteinsson. STRESS: super-resolution for dynamic fetal mri using self-supervised learning. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 197–206, 2021.

D. Xu et al. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10334–10343, 2019.

J. Xu et al. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7470–7482, 2023.

K. Xu et al. Masked modeling-based ultrasound image classification via self-supervised learning. *IEEE Open Journal of Engineering in Medicine and Biology*, 5:226–237, 2024.

B. Yaman et al. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic Resonance in Medicine*, 84(6):3172–3191, 2020.

K. Yan et al. SAM: self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Transactions on Medical Imaging*, 41:2658–2669, 2020.

B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye. Prototype mixture models for few-shot semantic segmentation. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pages 763–778, 2020.

D. Yang, J. Zhang, Y. Li, and Z. Ling. Skin lesion classification based on hybrid self-supervised pretext task. *International Journal of Imaging Systems and Technology*, 34(2), 2024.

J. Yang, D. Parikh, and D. Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5147–5156, 2016.

J. Yang, R. Shi, and B. Ni. MedMNIST classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195, 2021.

Q. Yang, W. Li, B. Li, and Y. Yuan. MRM: masked relation modeling for medical image pre-training with genetics. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21395–21405, 2023.

F. Yang et al. Self-supervised learning assisted diagnosis for mitral regurgitation severity classification based on color doppler echocardiography. *Annals of Translational Medicine*, 10(1):3, 2022.

J. Yang et al. Cross-modality segmentation by self-supervised semantic alignment in disentangled content space. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 52–61, 2020.

J. Yang et al. MedMNISTv2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023a.

X. Yang et al. Fetusmap: Fetal pose estimation in 3d ultrasound. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 281–289, 2019.

Y. Yang et al. Self-supervised interactive embedding for one-shot organ segmentation. *IEEE Transactions on Biomedical Engineering*, 70:2799–2808, 2023b.

M. Ye, X. Zhang, P. C. Yuen, and S. Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6203–6212, 2019.

C.-H. Yeh et al. Decoupled contrastive learning. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, pages 668–684, 2022.

P.-H. Yeung, A. I. L. Namburete, and W. Xie. Sli2Vol: annotate a 3d volume from a single slice with self-supervised learning. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 69–79, 2021.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 3320–3328, 2014.

C. You, R. Zhao, L. H. Staib, and J. S. Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 13434:639–652, 2021.

Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks, 2017.

K. Yu et al. DrasCLR: A self-supervised framework of learning disease-related and anatomy-specific representation for 3d lung ct images. *Medical Image Analysis*, 92:103062, 2024.

Y. Yuan, E. Ahn, D. Feng, M. Khadra, and J. Kim. SSPT-bpMRI: A self-supervised pre-training scheme for improving prostate cancer detection and diagnosis in bi-parametric MRI. In *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, volume 2023, pages 1–4, July 2023.

S. Yun, H. Lee, J. Kim, and J. Shin. Patch-level representation learning for self-supervised vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 8344–8353, 2022.

S. Yun et al. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019.

J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning, (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320, 2021.

X. Zhan, J. Xie, Z. Liu, Y. S. Ong, and C. C. Loy. Online deep clustering for unsupervised representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6687–6696, 2020.

B. Zhang, J. Xiao, and T. Qin. Self-guided and cross-guided learning for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8308–8317, 2021a.

C. Zhang, Y. Chen, L. Liu, Q. Liu, and X. Zhou. HiCo: Hierarchical contrastive learning for ultrasound video model pretraining. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 229–246, 2023a.

H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz. MixUp: beyond empirical risk minimization. In *6th International Conference on Learning Representations (ICLR) 2018*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=r1Ddp1-Rb`.

H. Zhang, S. Xu, W. Ren, H. Ye, and Y. Hong. Pretrain once and finetune many times: How pretraining benefits brain mri segmentation. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1724–1731, 2023b.

O. Zhang, M. Wu, J. Bayrooti, and N. D. Goodman. Temperature as uncertainty in contrastive learning. *arXiv*, abs/2110.04403, 2021b. URL `https://arxiv.org/abs/2110.04403`.

R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV), Part III 14*, pages 649–666, 2016.

S. Zhang, F. Zhu, J. Yan, R. Zhao, and X. Yang. Zero-CL: instance and feature decorrelation for negative-free symmetric contrastive learning. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*, 2022a. URL `https://openreview.net/forum?id=RAW9tCdVxLj`.

W. Zhang, J. Pang, K. Chen, and C. C. Loy. Dense siamese network for dense unsupervised learning. In *Proceedings of the European Conference on Computer Vision (ECCV), Part XXX*, pages 464–480, 2022b.

X. Zhang, Y. Wei, Y. Yang, and T. S. Huang. SG-One: similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50(9):3855–3865, 2020.

X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023c.

Y. Zhang, B. Hooi, D. Hu, J. Liang, and J. Feng. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 29848–29860, 2021c.

Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, volume 182 of *Proceedings of Machine Learning Research*, pages 2–25, 2022c.

H. Zhang et al. JIANet: jigsaw-invariant self-supervised learning of autoencoder-based reconstruction for melanoma segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–13, 2022a.

K. Zhang et al. Guided networks for few-shot image segmentation and fully connected CRFs. *Electronics*, 9(9), 2020.

S. Zhang et al. Align representations with base: A new approach to self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16579–16588, 2022b.

X. Zhang et al. Self-supervised tumor segmentation with Sim2Real adaptation. *IEEE Journal of Biomedical and Health Informatics*, 27:4373–4384, 2023a.

Y. Zhang et al. A point in the right direction: Vector prediction for spatially-aware self-supervised volumetric representation learning. In *Proceedings of the 20th IEEE International Symposium on Biomedical Imaging, ISBI*, pages 1–5, 2023b.

Y. Zhang et al. Geometric view of soft decorrelation in self-supervised learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 4338–4349, 2024a.

Y. Zhang et al. A self-supervised fusion network for carotid plaque ultrasound image classification. *Mathematical Biosciences and Engineering*, 21(2):3110–3128, 2024b.

C. Zhao, A. Carass, B. E. Dewey, and J. L. Prince. Self super-resolution for magnetic resonance images using deep networks. In *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, pages 365–368, 2018.

C. Zhao et al. SMORE: a self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning. *IEEE Transactions on Medical Imaging*, 40(3):805–817, 2021.

L. Zhao et al. Medical image segmentation based on self-supervised hybrid fusion network. *Frontiers in Oncology*, 13:1109786, Apr. 2023.

M. Zheng et al. ReSSL: relational self-supervised learning with weak augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 2543–2555, 2021.

R. Zheng et al. MsVRL: self-supervised multiscale visual representation learning via cross-level consistency for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42:91–102, 2022.

Y. Zhi, H. Bie, J. Wang, and L. Ren. Masked autoencoders with generalizable self-distillation for skin lesion segmentation. *Medical & Biological Engineering & Computing*, 2024.

H.-Y. Zhou, C. Lu, S. Yang, X. Han, and Y. Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3479–3489, 2021a.

H.-Y. Zhou, C.-K. Lu, C. Chen, S. Yang, and Y. Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:8020–8035, 2023.

Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021b.

Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition, 2012.

J. Zhou et al. Image BERT pre-training with online tokenizer. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022a. URL `https://openreview.net/forum?id=ydopy-e6Dg`.

P. Zhou et al. Mugs: A multi-granular self-supervised learning framework. *ArXiv*, abs/2203.14415, 2022b.

Y. Zhou et al. A simple framework uniting visual in-context learning with masked image modeling to improve ultrasound segmentation. *arXiv preprint arXiv:2402.14300*, 2024.

C. Zhu, W. Chen, T. Peng, Y. Wang, and M. Jin. Hard sample aware noise robust learning for histopathology image classification. *IEEE Transactions on Medical Imaging*, 41(4):881–894, 2022.

K. Zhu, W. Zhai, Z. Zha, and Y. Cao. Self-supervised tuning for few-shot segmentation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1019–1025, 2020.

R. Zhu, B. Zhao, J. Liu, Z. Sun, and C. W. Chen. Improving contrastive learning by visualizing feature transformation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10286–10295, 2021.

J. Zhu et al. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis*, 64:101746, 2020.

J. Zhu et al. TiCo: transformation invariance and covariance contrast for self-supervised visual representation learning. *CoRR*, abs/2206.10698, 2022. URL `https://doi.org/10.48550/arXiv.2206.10698`.

C. Zhuang, A. Zhai, and D. Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6001–6011, 2019.

X. Zhuang et al. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 420–428, 2019.

D. Ziegler et al. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL `http://arxiv.org/abs/1909.08593`.

Álvaro S. Hervella, J. Rouco, J. Novo, and M. Ortega. Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction. *Applied Soft Computing*, 91: 106210, 2020.

Álvaro S. Hervella, J. Rouco, J. Novo, and M. Ortega. Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis. *Expert Systems with Applications*, 185:115598, 2021.