Environmental Degradations in Images: Analysis, Restoration, and Applications

Harsh Bhandari

Computer Vision and Pattern Recognition Unit (CVPRU)
Indian Statistical Institute
203 B. T. Road
Kolkata - 700 108, India

Supervisor: Prof. Sarbani Palit



A thesis submitted to the Indian Statistical Institute
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science
April 2025

To my family...

Acknowledgements

This thesis is the perfect platform to thank numerous people, including my parents, relatives, teachers, friends, colleagues, and all well-wishers whose immense, inspiring support and encouragement during the research work have driven me on this interesting and challenging journey. I gleefully grab this opportunity to extend my sincere gratitude and appreciation to all whose constant support, contribution, and guidance, made the smooth completion of my Ph.D. thesis. , I am immensely indebted to my supervisor, Prof. Sarbani Palit, for her kind and invaluable supervision. Her motivation, suggestions, wisdom, and crucial remarks encouraged and boosted my confidence and belief in the completion of my thesis. I would like to thank her for thoroughly checking and valuable suggestions, which have contributed greatly to the improvement of the thesis.

Besides my advisor, I would like to thank the Junior Research Fellow (JRF) selection committee members of the Indian Statistical Institute (ISI) for selecting me as a research scholar at this Institute. I also like to thank the Computer and Communication Sciences Division - Research Fellow Advisory Committee (CCSD-RFAC) members for their great support, constructive suggestions, and invaluable advice at every evaluation meeting throughout my research.

I am indebted to my colleagues and friends for their useful discussions and friendly advice. I would also like to thank all the faculty members and the office staff of the Computer Vision and Pattern Recognition Unit (CVPRU) and other Units of ISI for their support.

Finally, I am very grateful to my family for their unconditional love, continued patience, and endless support, without which this journey would not have been possible. In conclusion, I would like to thank all well-wishers (including those whom I might have missed here) who are connected with me (directly or indirectly) in this fruitful journey. Thank you all.

Harsh Bhandari

Abstract

When dealing with real-world images and video taken in an outdoor environment, we often encounter scenes effected by environmental degradations such as those impacted by haze, fog or underwater. Many real-world applications, such as autonomous driving, surveillance, underwater archaeology, and marine biology suffer due to such degradations. A common approach to deal with such degradation is to judiciously clean the images following image enhancement or image restoration approaches, and use the clear data for further processing. The objective of this thesis is to develop methodologies to handle the image degradation due to haze and underwater effects using computational intelligence and data-driven deep learning methods. To develop such methods we normally use a set of labeled datasets. However, one can also use a mixture of labeled and unlabeled datasets. We emphasize that our goal is beyond image restoration. Handling of environmentally degraded images by enhancement or restoration is just one facet. We also use latent space representation for dealing with other applications of degraded images, i.e., estimating the concentration of particulate matter in air, which is completely different from image restoration though effective image restoration does serve as an intermediary objective in this case.

We now discuss the contribution of this thesis briefly, chapter-wise.

In the first contributory chapter, we have proposed a pyramid fusion approach to restore a hazy image. The proposed approach to perform dehazing neither uses any pre-trained model based on some benchmark dataset nor tries to compute the standard haze parameters - transmittance and air light. The method extracts image-based features from the image frame(s) that are highly correlated with the haze. Since the presence of haze may be non-uniform in the frame, the proposed approach extracts features from patches to capture the heterogeneity. Further, the method uses a multi-scale approach to extract features from different scales that are used in the creation of a pyramid. Finally, the model uses Gaussian and Laplacian filters to remove the impact of haze and fuse the refined patches to generate a clear haze-free image.

Apart from the hazy images restoration, the thesis focuses on its application. Two major applications are an image-based multi-weather detection system and estimating the concentration of particulate matter present in surrounding air using an image-

based approach. Very few research works have been conducted in this area due to the availability of limited data. A robust dataset has been created such that the images capture the concentration of the particulate matter at different intervals of time throughout the day and night. It also covers the variation in the concentration on account of seasonal changes over a year. The dataset includes more than 60000 images covering the day and night parts separately. Measuring the particulate concentration requires two major pieces of information apart from images - season and depth of the surrounding area. Our next two chapters will focus on determining the weather of the surrounding area as well as depth mapping from the camera to the nearby surroundings.

In the second contributory chapter, a novel image-based multi-weather detection model has been proposed. The image is segmented into small clusters using the Mean-Shift clustering method, which are then used to extract features. Sky plays a vital role in judging the weather condition of the surrounding, hence a U-net based model has been designed to extract the sky region from the image. Features extracted from sky and non-sky regions are fused to create an 84-D feature vector that is fed to the modified Support Vector Machine(SVM) to predict the weather condition observed in an image.

The second contributory chapter also estimates the concentration of particulate matter of size less than 2.5 μm present in outdoor air from daytime images. The method utilizes the dehazed version of the daytime image, the structural similarity feature, and the date and time of the image's capture. The model considers the full image and region of interests (RoIs) for feature extraction. Considering the estimation as a regression problem, all the extracted features are provided as input to multiple machine learning algorithms to compute the particulate concentration and look for the best model that provides the most accurate values.

The third contributory chapter deals with estimating the concentration of particulate matter from nighttime images. Night images used for particulate estimation suffer from two major challenges - the spectral reflectance of light by the particulate matters and the distance of various structures and objects from the camera. Depth information plays an important role in accurately estimating the concentration of particulate matter in air. The visibility of an area is inversely proportional to the depth and decreases exponentially with it. Thus, estimating the concentration in the absence of depth information may lead to miscalculation as high concentration at the near region can have similar features with the farther low visible regions with low concentration. This necessitates the development of a depth map from an image of night scenes. The

depth map is based on a deep-learning model trained on labeled benchmark datasets from night images. The fusion of features generated from depth maps and original images has been used to train our encoder-decoder model to estimate the concentration of particulate matter in the region covered in the image. Thus, this chapter covers two aspects of the thesis. The first is depth estimation from both day and night images, and the second is the utilization of the model in estimating the concentration level from night scenes.

Moving from open-air scenes, the fourth contributory chapter dives into the underwater scenes. Underwater scenes suffer from some major drawbacks. As the distance from the camera to the underwater objects increases, the frames suffer from major color loss starting with the red wavelength followed by green. The presence of water molecules also adds blurriness to the image caused by the refraction of light from the large water particles. A novel deep learning model based on a transformer has been proposed with some major modifications made to the Swin Transformer to suit the vision application.

Contents

Contents xi

List of Figures xv

List of Tables xix

List of Algorithms xxi

1 Introduction:Scope of the Thesis 1

1.1 Introduction . . . 1

1.1.1 Image Dehazing . . . 3

1.1.2 Encoder-Decoder network . . . 3

1.1.3 Transformer . . . 4

1.1.4 Underwater-image Restoration . . . 5

1.1.5 Conditional Random Fields(CRF) . . . 6

1.1.6 Particulate Matter Concentration . . . 7

1.1.7 Monocular Depth Estimation . . . 8

1.1.8 Particulate Matter Estimation . . . 8

1.1.9 Datasets . . . 9

1.2 The Scope of the Thesis . . . 10

1.3 Conclusions, Limitations, and Future Scopes . . . 11

1.4 The major contributions of the thesis . . . 12

2 Image Dehazing: Pyramid Fusion Approach 15

2.1 Introduction . . . 15

2.2 Related Work . . . 17

2.2.1 Traditional approach . . . 17

2.2.2 Multi-result fusion approach . . . 18

2.2.3 Multi-feature fusion approach 18
2.3 Proposed Methodology 19
2.3.1 Formulation of the approach 19
2.3.2 Feature Extraction 23
2.3.3 Image Fusion 26
2.4 Experiments 29
2.4.1 Datasets, Parameter Settings & Simulation Environment 29
2.4.2 Performance Analysis 32
2.4.3 Limitations 32
2.5 Concluding remarks 32

3 Application of Machine Learning using image-based features for Multi-Weather Detection and Particulate Matter Concentration Estimation 33
3.1 Introduction 33
3.2 Image based Multi-weather detection 35
3.2.1 Related Work 35
3.2.2 Weather Features 38
3.2.3 CNN Model 48
3.2.4 Datasets and MLJSVM 48
3.2.5 Analysis of experiments and results 51
3.3 Image-based Determination of Particulate Matter Concentration 55
3.3.1 Related Work 56
3.3.2 Formulation of the approach 56
3.3.3 Experimental results for Particulate Matter Concentration system 60
3.3.4 Concluding remarks 64

4 Estimation of Particulate Matter Concentration from night scenes: From REAL-Net to Dark-AirNet 69
4.1 Introduction 69
4.2 Related Work 71
4.2.1 **Machine Learning based PM concentration estimation** 71
4.2.2 **Deep Learning based concentration estimation** 72
4.3 Dataset 74
4.4 The Proposed Method 75
4.4.1 High-Low Frequency Block 75
4.4.2 Depth-Structural Module (REAL-Net) 78
4.4.3 Residual Attention Block 82

4.4.4 Multi-scale Fusion . . . 84
4.5 Experiment and Evaluation . . . 84
4.5.1 Training Set . . . 84
4.5.2 Loss function . . . 85
4.5.3 Implementation details for REAL-Net . . . 86
4.5.4 Comparison methods for depth estimation model . . . 86
4.5.5 Ablation Study on REAL-Net . . . 87
4.5.6 Evaluation Metrics . . . 87
4.5.7 Observed Patterns in Particulate Matter Concentration . . . 89
4.5.8 Annual Observation . . . 90
4.5.9 Observation during the rainy season . . . 90
4.5.10 Observation during the autumn season . . . 91
4.5.11 Observation during the winter season . . . 91
4.5.12 Model Settings . . . 91
4.5.13 Results: Performance comparison and quantitative assessment . . . 92
4.5.14 Analysis of Results and Ablation Study . . . 98
4.6 Concluding Remarks . . . 100

5 Aquaformer: Underwater Image Enhancement via Adaptive Transformer 103
5.1 Introduction . . . 103
5.2 Related Work . . . 104
5.2.1 Physical Models . . . 104
5.2.2 Visual Prior Models . . . 105
5.2.3 Data-driven Models . . . 105
5.3 Proposed Model . . . 106
5.3.1 An overview . . . 106
5.3.2 Statistical Adaptive Layer Normalization . . . 107
5.3.3 Symmetric Padding on Shifted Window . . . 108
5.3.4 Adaptive Fusion . . . 109
5.3.5 Convolved W-MHSA mechanism . . . 109
5.4 Implementation Details with Experimental Data . . . 110
5.4.1 Implementation Details . . . 111
5.4.2 Datasets . . . 112
5.4.3 Evaluation Metrics . . . 113
5.4.4 Comparison of performance . . . 114
5.4.5 Color Restoration Performance . . . 114

5.4.6 Visual Quality Performance 115
5.4.7 Ablation Study 116
5.5 Concluding Remarks 116

6 Conclusions **121**
6.1 A Summary of the work done 121
6.2 Scope for Future Work 124

Bibliography **127**

List of Figures

1.1 Architecture of Transformer 5

1.2 Underwater-image-formation-model 6

1.3 Sample images from multiple benchmark datasets 9

2.1 Image Dehazing using the proposed algorithm 16

2.2 Contrast Enhancement: (a) Original Image, (b) HE, (c) CLAHE (d) Proposed 20

2.3 The block diagram of the proposed approach 21

2.4 The overall process of the proposed algorithm is illustrated using a hazy image as an example. In the first column, the input hazy image is displayed. The second column contains the preprocessed images, with the top image being White-Balanced and the bottom one showing the Contrast Enhanced version. Columns three through seven present the feature maps for the corresponding derived images, representing Chromaticity, Colorfulness, Entropy, Saliency, and Visibility, respectively. The eighth column shows the normalized feature map generated from each input-derived image. Finally, the dehazed output is obtained by applying the fusion process to the two resultant images, as shown in Fig. 2.1(b). 23

2.5 Variation in the Saliency and Colorfulness Map using the CHIC Database [1] with the haze level increasing from right to left, where the last column represents the ground truth 27

2.6 The block diagram of the Pyramid Fusion approach 28

2.7 A comparison of single image dehazing algorithms for heavily hazed images is conducted using the O-haze Database[2] and the I-haze Database[3] 30

3.1 The proposed weather detection method includes two main components: (A) the SKYNET model, which generates sky and non-sky segments from an input image, and (B) the advanced ResNet-101 model, which extracts machine-based features. Handcrafted features are extracted from the segmented sky and non-sky regions and combined with the ResNet features. The combined features are fed into the MLJSVM classifier to generate the final output. 35
3.2 The Five weather classes - Rainy, Cloudy, Sunny, Haze, and Snow are listed from left to right, respectively. 38
3.3 Highlighting the impact of illumination in detecting sunlight-bearing segments. 40
3.4 Colorfulness of images (a) Sunny (b) Cloudy (c) Rainy (d) Snow (e) Hazy 41
3.5 The DCP maps are displayed in the second row, with the first three columns showing images with white haze and the last three columns showing images with non-white haze. 42
3.6 Scenes with clear and natural haze are presented, where the top row contains both hazy and haze-free images, and the bottom row shows their corresponding Visibility Maps. 43
3.7 Sky Segmentation - SKYNET 45
3.8 Sky detection results for (a) cloudy image, (b) rainy image, (c) sunny image, (d) snowy image, and (e) hazy image. 48
3.9 Sky Mask for (a) cloudy image (b) rainy image (c) sunny image (d) snow image (e) hazy image 50
3.10 Ablation study of segment numbers with respect to the extracted feature. 52
3.11 Block diagram of the proposed method. 57
3.12 Visibility of selected RoIs with respect to $PM_{2.5}$ concentration. 58
3.13 Selection of RoIs based on the relative distance of objects (one near and one relatively far) 59
3.14 Visibility and SSIM of selected RoIs after dehazing. 60
3.15 Images along with dehazed image and SSIM feature map of low and high $PM_{2.5}$ respectively. 61
3.16 Images and their corresponding SSIM feature map from the multi-view dataset. 62
3.17 Variation of $PM_{2.5}$ concentration from Jul-2021 to Jan-2023. 63
3.18 Distribution Map 66

3.19 Selection of RoIs based on the relative distance of objects (one near and one relatively far) on Shanghai images. 67
3.20 Scatter plots (Estimated $PM_{2.5}$ vs Predicted $PM_{2.5}$) of RoI-based Models a) SVR, b) RFR, and c) ETR, and entire image-based Models d) ETR on Single-View dataset, e) ETR model on Multi-View Dataset. 68

4.1 Architecture of Dark-AIRNET 76
4.2 Architecture of High-Low Frequency block extracting the frequency-aware features. 77
4.3 Architecture of Monochromatic Depth Estimation module extracting the depth-based regional information from the encoder. The model utilizes the extracted attention-encoded features and spatial-feature value attention layer in the decoder to produce the estimated depth map. The deep blue arrows represent the flow of attention-guided information. . . 79
4.4 Depth map presentation of night images with the top row indicating color images and the bottom row with their corresponding depth maps. 82
4.5 Architecture of Residual Attention Block(RAB). 83
4.6 Visual comparison of depth estimation on images from the KITTI [4] dataset. The first to the fifth columns correspond to the color image and the results obtained from [5], [6], [7], and our proposed method, respectively. 87
4.7 The visual comparison of depth estimation on images from the Cityscapes [8] dataset is shown. The first column presents the color image, followed by the depth results obtained from [6], [7], and our proposed method in the second, third, and fourth columns, respectively. 88
4.8 Scatter plots of Measured vs Estimated PM in order - (a) Annual, (b) Autumn, (c) Rainy, and (d) Winter respectively. 90
4.9 Pattern exhibited in $PM_{2.5}$ data collected throughout the year between July 2021 and April 2022 92
4.10 Pattern exhibited in PM_{10} data collected throughout the year between July 2021 and April 2022 92
4.11 Pattern exhibited in $PM_{2.5}$ data collected during the period of July-September, 2021 93
4.12 Pattern exhibited in PM_{10} data collected during the period of July-September, 2021 93
4.13 Pattern exhibited in $PM_{2.5}$ data collected during the period of October-December, 2021 93

4.14 Pattern exhibited in PM_{10} data collected during the period of October-December, 2021 94
4.15 Pattern exhibited in $PM_{2.5}$ data collected during the period of January-March, 2022 94
4.16 Pattern exhibited in PM_{10} data collected during the period of January-March, 2022 94
4.17 Seasonal error distribution for (a) $PM_{2.5}$ and (b) PM_{10} respectively. . . . 97
4.18 Images taken on 02-Dec-2021, at 07:02, 09:04, 09:49, and 10:44 PM respectively. 101
4.19 Images having structural variations taken on 01-Apr-2022, at 10:41, 10:46, 10:48, and 10:51 PM respectively. 101
4.20 Images taken during summer(1st Apr 2022), rainy(6th Dec 2021), autumn(3rd Nov 2021), and winter(1st Feb 2022) respectively. 102
4.21 Images captured under different light conditions on 3rd Nov 2021, 6th Oct 2021, 2nd Feb 2022 and 10th Feb 2022 respectively. 102

5.1 Architecture of a 5-stage Aquaformer with a detailed presentation of an Aquaformer Block 107
5.2 Adaptive fusion approach 110
5.3 Visual quality performance comparison on images (row-wise) with the first three from Test-R90[9] followed by the next three from SUIM-E[10], two from EUVP[11], and the last two from Test-S1000[12] test dataset. Zoom to the maximum extent possible to get full clarity. 117
5.4 Visual quality performance comparison (row-wise images) with the first three images from Test-C60[9], the next three from SQUID[13], and the last two image-colorchecker pair from Color-Checker7[14] dataset. Zoom to the maximum extent possible to get full clarity. The smaller the CIEDE score, the better the result. 118

List of Tables

2.1 Comparison of Dehazed Images with AMEF[15], CAP[16], and, BCCR[17] shown in Fig. 2.7 . . . 31
2.2 Comparison of Dehazed Images with DNet[18] and Defade[19] models shown in Fig. 2.7 . . . 31

3.1 Sky Properties . . . 45
3.2 Classification results obtained using only the specified features are presented . . . 53
3.3 Classification results obtained by omitting the specified feature and the features extracted from ResNet-101 are presented with all figures in %. . 53
3.4 Classification results of different methods. . . . 53
3.5 Results of RoI-Based Models as per Dehazing Methods: Pyramid Fusion(PF) and Colour Attenuation Prior(CAP). . . . 62
3.6 Results of Entire Image-Based Models based on ETR for different Dehazing Methods: Pyramid Fusion(PF) and Colour Attenuation Prior(CAP). 63
3.7 Comparison of performance of the proposed methods with PPPC model on Single-View, Multi-View, and Shanghai Datasets. . . . 64

4.1 Ablation Study on the KITTI dataset. . . . 87
4.2 Quantitative visual quality performance on KITTI dataset . . . 88
4.3 Quantitative visual quality performance on Cityscapes dataset. . . . 88
4.4 Quantitative visual quality performance on Foggy Zurich dataset. . . . 89
4.5 Ablation Study results of the network. The best result for each metric has been displayed in bold font. . . . 89
4.6 AQI category, pollutants, and health breakpoints in $\mu g/m_3$ . . . 89
4.7 Quantitative performance achieved by models for particulate estimation. The best result for each metric has been displayed in bold font. . . 95
4.8 Performance Measures of Dark-AIRNET . . . 96

4.9 Estimated $PM_{2.5}$ predicted by the proposed model Dark-AirNet along with their actual value for images shown in the respective figure numbers in left to right format respectively. 99
4.10 Estimated $PM_{2.5}$ predicted by Resnet-50 model along with their actual value for images shown in the respective figure numbers in left to right format respectively. 99
4.11 Estimated PM_{10} predicted by the proposed model Dark-AirNet along with their actual value for images shown in the respective figure numbers in left to right format respectively. 100
4.12 Estimated PM_{10} predicted by Resnet-50 model along with their actual value for images shown in the respective figure numbers in left to right format respectively. 100

5.1 Ablation Study for judging the effectiveness of the contributions, conducted on Test-R90 and SUIM-E 116
5.2 Quantitative comparison among SOTA methods using reference testing set. The top three scores of PSNR(P) and SSIM(S) in descending order are marked with red, green, and blue respectively 119
5.3 Qualitative comparison among SOTA methods. The top three scores of BRISQUE(B), NIQE(N), and PIQE(P) in ascending order are marked with red, green, and blue respectively 119

List of Algorithms

1 Contrast Enhancement 22

Chapter 1

Introduction:Scope of the Thesis

> A lot of life is about trying to turn bad experiences into something good. Usually if you work at it, you can figure out a way to do it. Even our worst misfortunes are gifts. –Robert Kennedy, Jr.
>
> It's not that I'm so smart, it's just that I stay with problems longer.– Albert Einstein

1.1 Introduction

The world we inhabit is characterized by a rich diversity of objects—both natural and human-made—each contributing to a vibrant palette of colors and forms. The sky itself reflects this variety. However, environmental conditions, whether naturally occurring or anthropogenic, often interfere with how we visually perceive our surroundings. Among such conditions, fog and haze are particularly common, diminishing the clarity of shapes and colors in a scene. The extent of this visual degradation depends on the severity of the phenomenon. While a light mist might enhance the aesthetic quality of a photograph, dense fog significantly reduces visibility, posing serious challenges for activities like driving and commuting.

In regions like northern and eastern India, winter fog causes substantial disruption to railway schedules, leading to delays and inconveniences for travelers. Road safety is also compromised, with increased occurrences of accidents during foggy conditions. Capturing images of such low-visibility scenes using standard RGB cameras and subsequently applying dehazing techniques can restore image sharpness and color fidelity. This enhancement is especially beneficial for driver assistance technologies and autonomous navigation systems, where clear vision is essential but still difficult

to achieve in hazy conditions, even with sophisticated sensors like LIDAR. Furthermore, dehazing can support forensic investigations by making crucial details, such as license plate numbers, visible in poor weather.

Beyond restoration, visibility loss due to weather can also serve as an informative signal. For instance, classifying weather types using visual cues from images opens up new applications in traffic safety, autonomous vehicle systems, and image preprocessing pipelines. Recognizing specific weather patterns from images taken in the past can aid forensic reconstructions and help analyze incidents more effectively.

A major contributor to reduced visibility and deteriorated air quality is particulate matter (PM), particularly those particles under 2.5 microns ($PM_{2.5}$), which can bypass nasal filters and penetrate deep into the lungs, causing respiratory issues. Monitoring PM_{10} and $PM_{2.5}$ levels is crucial for assessing air quality and is commonly performed using dedicated monitoring stations. However, these setups are expensive and infrastructure-heavy, prompting the need for affordable, image-based alternatives.

Since particulate pollution directly impacts how light travels through the air, leading to haziness, this characteristic can be leveraged to develop a vision-based estimation system for $PM_{2.5}$ levels. By using high-resolution images that adequately capture outdoor scenes, it may be possible to infer pollutant concentrations in a given area. However, several environmental variables complicate this task, such as light rain, mist, and seasonal fog, all of which affect visibility and must be factored into prediction models. A large dataset of images with corresponding ground-truth pollution levels is essential to train and evaluate such systems.

The complexity increases dramatically at night, where natural light is absent, and artificial illumination introduces its own challenges. Nighttime images suffer from noise, inconsistent lighting, and color distortions depending on light sources, all of which hinder visibility-based pollution estimation. Overcoming these obstacles requires systems specifically tailored to nighttime imaging, including robust methods for denoising and handling spectral interference.

A critical piece of information for accurately estimating PM levels is the distance between the camera and the scene objects. Estimating this from images—especially at night—is notoriously difficult due to fog, varied lighting, and other environmental factors. A promising direction is to develop separate models for day and nighttime depth estimation, and then integrate depth maps into the PM prediction pipeline for improved accuracy.

Interestingly, similar visual degradation issues occur underwater, where light behaves differently. Water absorbs light frequencies unevenly, with red light fading

quickly and blue light penetrating deeper. This causes color distortion and blurriness in underwater images, especially as depth increases. Other factors like water turbidity, particle suspension, and movement-induced disturbances further impair image quality. Enhancing underwater images is crucial for tasks such as marine exploration, where visibility is key to navigation and object recognition. Pairing underwater imagery with real-time water quality data can also open doors to image-based aquatic monitoring systems.

A brief piece of information regarding each of the subjects covered in the thesis is discussed herewith.

1.1.1 Image Dehazing

Initiated upon the foundational principles of the atmospheric scattering model (ASM) [20], early endeavors in image dehazing have been primarily oriented towards delineating the correlation between hazy images and their haze-free counterparts. The ASM can be succinctly articulated as per the Eq. (1.1.1):

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1.1.1}$$

where I and J signify the hazy image and its clean counterpart; x and A represents the pixel location and the global atmosphere light; t(x) denotes the transmission map, which is a function of the atmosphere scattering parameter β and the scene depth $d(x)$, articulated as $t(x) = e^{-\beta d(x)}$. This formulation posits the image dehazing to the precise estimation of the transmission map t(x) and the global atmosphere light A. Performing the inverse operation on the equation to generate J(x) recreates the clear, haze-free image.

1.1.2 Encoder-Decoder network

The encoder-decoder architecture is a neural network framework commonly used in sequence-to-sequence tasks. It comprises two main components: an encoder and a decoder. The encoder's role is to process an input sequence and transform it into a set of context vectors that encapsulate the essential information of the input. These context vectors are then passed to the decoder, which uses them to produce a corresponding output sequence. This setup is particularly useful for tasks like machine translation, text summarization, and image caption generation, where one type of data needs to be converted into another. The core concept behind this architecture is to enable machines to learn the complex mappings between different data modalities. The

encoder captures the essence of the input, while the decoder attempts to reconstruct or generate the appropriate output based on this encoded information. For accurate reconstruction or generation, there must be a meaningful correlation between the input and the desired output—without this relationship, the model's output would be arbitrary. In our chapters dealing with underwater-image restoration, depth estimation, and particulate concentration estimation, we use an encoder-decoder neural network architecture trained on multiple benchmark datasets.

Next, we very briefly discuss the concepts of transformer, image-dehazing, and underwater image restoration so that the subsequent discussion on the literature becomes easy to follow. Moreover, these concepts will be required for the development of the theories and methodologies in the subsequent chapters of this thesis.

1.1.3 Transformer

A transformer model [21] as shown in Fig. 1.1 is a neural network that learns the context of sequential data and generates new data out of it. Originally devised for sequence transduction or neural machine translation, transformers excel in converting input sequences into output sequences. It is the first transduction model relying entirely on self-attention to compute representations of its input and output without using a sequence-aligned Recurrent-Neural Network or a convolution. The main core characteristic of the Transformers architecture is that it maintains the encoder-decoder model. The input image is of type $\mathbb{R}^{H\times W\times C}$, where H, W, C are height, width, and channel (RGB), respectively. Then it is divided into square patches of type $\mathbb{R}^{P\times P\times C}$. In this approach, each image patch is first passed through a linear projection layer to generate a corresponding vector, often referred to as a patch embedding. Simultaneously, the spatial location of each patch is encoded into a position embedding vector. These two vectors—the patch embedding and the positional encoding—are then combined and fed into a stack of Transformer encoder layers. Within the Transformer, the attention mechanism progressively refines these patch representations by modeling increasingly complex semantic relationships among them. This process mirrors how Transformers in natural language processing enhance word representations as they propagate through layers, progressing from basic syntactic understanding to deeper semantic comprehension. By treating the image as a sequence of embedded patches, this architecture effectively transforms it into a series of meaningful vector representations. To utilize these representations for specific tasks (such as classification, segmentation, or detection), a task-specific head must be trained on top of the transformer output. Transformers are excellent in capturing long-range dependencies

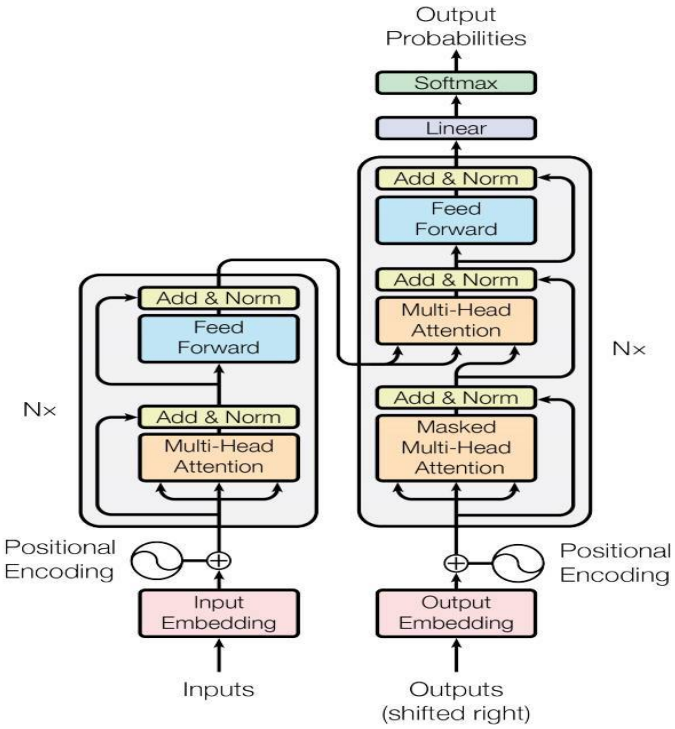


Figure 1.1: Architecture of Transformer

that lie in image data. This makes the model more robust compared to those based on convolution-based architectures. In the chapter dealing with underwater image restoration, our model is based on a transformer architecture responsible for color correction and image enhancement.

1.1.4 Underwater-image Restoration

In the ocean, inherent and apparent optical properties (IOPs and AOPs) govern light propagation, with Fig. 1.2 demonstrating a simpler underwater image formation model. The IOPs are only a function of the water constituents. These are beam absorption a(λ), beam scattering b(λ), and beam attenuation $\beta(\lambda)$ coefficients, where a(λ)+b(λ) = $\beta(\lambda)$. The AOPs depend on external factors such as the ambient light field [22], but are easier to measure than IOPs. Here, the only AOP we use is Kd(λ), the diffuse attenuation coefficient of the downwelling light, which is the extinction experienced by light penetrating a water column vertically. The magnitude of Kd only weakly depends on ambient light, so it has been used to characterize the optical qual-

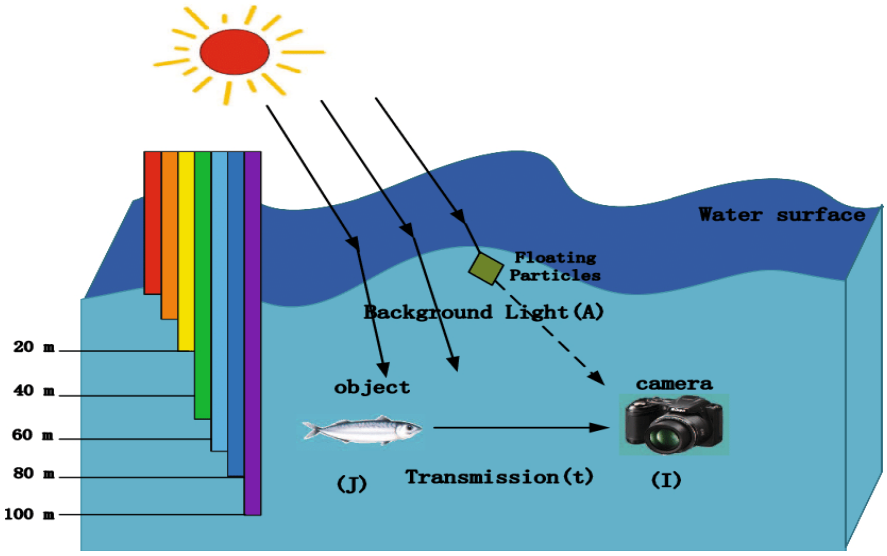


Figure 1.2: Underwater-image-formation-model

ities of water bodies, such as Jerlov's water types[23]. Based on findings made in [24], the underwater image formation model can be described as per the Eq. (5.3.3):

$$I_c = J_c e^{-\beta_c^D(v_D).z} + B_c^{\infty}(1 - e^{-\beta_c^B(v_B).z}) \tag{1.1.2}$$

Here, I_c and J_c refers to the underwater degraded and clear image at color channel c = {R, G, B} respectively, β_c^D = wideband attenuation coefficients derived from the direct transmission (D), the vectors v_D and v_B represent the coefficient dependencies v_D = {z, ρ, E, S_c, β} and v_B = {E, S_c, b, β}, z is the horizontal distance of the camera sensors from the object, B^{∞} = backscatter at infinity, also termed veiling light, E(d,λ) is ambient light at depth d, $\rho(\lambda)$ = scene reflectance, $b(\lambda)$ = beam scattering, $S_c(\lambda)$ = spectral response for color channel c at wavelenght λ. Similar to image dehazing, the inverse operation performed on the equation [5.3.3] leads to a clear underwater enhanced image.

1.1.5 Conditional Random Fields(CRF)

Conditional Random Fields is a class of discriminative models best suited to prediction tasks where contextual information or the state of the neighbors affect the current prediction. CRF is a special case of Markov Random field wherein the graph satisfies the property: "When we condition the graph on X globally i.e. when the values of random variables in X is fixed or given, all the random variables in set Y follow the

Markov property p(Y_u/X,Y_v, $u \neq v$) = p(Y_u/X,Y_x, $Y_u \sim Y_x$), where $Y_u \sim Y_x$ signifies that Y_u and Y_x are neighbors in the graph." In our chapter dealing with depth estimation, we have fused the concept of CRF with deep learning, which fuses complementary information derived from multiple CNN side outputs. By designing a novel CNN implementation of mean-field updates for continuous CRFs, we show the effectiveness of the proposed model in estimating the depth information from a color RGB image.

1.1.6 Particulate Matter Concentration

Airborne particulate matter (PM) is not a single substance, but a diverse blend of various chemical elements. It consists of a combination of solid particles and liquid droplets, including dry fragments, small liquid drops, and solids coated with liquids. These particles can differ significantly in size, shape, and composition, and may include substances such as inorganic ions, metals, elemental carbon, organic chemicals, and materials originating from the Earth's crust. For regulatory and air quality monitoring purposes, particulate matter is categorized based on its diameter. PM_{10} refers to particles with diameters of 10 micrometers or smaller, which are small enough to be inhaled into the respiratory system and potentially cause negative health effects. A subset of this group, known as $PM_{2.5}$, includes particles that are 2.5 micrometers or smaller. Since $PM_{2.5}$ particles fall within the size range of PM_{10}, they are considered a component of the broader PM_{10} category.

$PM_{2.5}$

$PM_{2.5}$ refers to particulate matter with a diameter of 2.5 micrometers or smaller, which is about 3% the diameter of a human hair. These tiny particles are primarily produced by combustion processes, such as those from vehicles, industrial facilities, and residential heating, as well as agricultural burning and wildfires. They can also form secondarily from gases like sulfur dioxide and nitrogen oxides. Health-wise, $PM_{2.5}$ particles are particularly concerning because they can penetrate deep into the lungs and enter the bloodstream. This can lead to respiratory issues like asthma and bronchitis, cardiovascular problems such as heart attacks and strokes, and even premature death in individuals with existing heart or lung conditions. Environmentally, $PM_{2.5}$ reduces visibility, contributing to haze, and can settle on the ground or water, leading to environmental degradation and the formation of acid rain.

PM_{10}

PM_{10} refers to particulate matter with a diameter of 10 micrometers or smaller. These larger particles can be inhaled and originate from sources like road and construction dust, industrial activities, pollen, mold spores, sea spray, and combustion processes. Health effects of PM_{10} include penetration into the respiratory tract and reaching the lungs, causing respiratory issues such as coughing, wheezing, and shortness of breath, and aggravating existing respiratory and cardiovascular diseases. While they are less likely than $PM_{2.5}$ to penetrate deep into the lungs, they still pose significant health risks. Environmentally, PM_{10} reduces visibility and can settle on surfaces, leading to soil and water pollution. These particles can also damage crops and forests by interfering with photosynthesis.

1.1.7 Monocular Depth Estimation

Monocular Depth Estimation is the task of estimating the depth value (distance relative to the camera) of each pixel given a single (monocular) RGB image. This challenging task is a key prerequisite for determining scene understanding for applications such as 3D scene reconstruction, autonomous driving, and augmented reality (AR). In the thesis, monocular depth estimation has been applied to estimate the concentration of particulate matter accurately from a given nighttime image. Depth information helps the model in discriminating areas with poor visibility due to either large depth or because of a high concentration of particulate matter.

1.1.8 Particulate Matter Estimation

Numerous methods exist for estimating the concentration of particulate matter. To estimate the surface $PM_{2.5}$ and PM_{10} concentrations, different linear and nonlinear multivariable regression models are available in the literature. Here we discuss the Beta attenuation process, as this is the golden method used by the Government of India to estimate the concentration levels. When beta particles interact with a material, they may be absorbed, reflected, or transmitted through it. The degree to which beta radiation is reduced, or attenuated, is directly related to the amount and type of material present. Beta attenuation monitors, also known as beta gauges, operate on the principle that beta particles lose energy when they pass through particulate matter (PM) collected on a filter. To isolate the attenuation caused solely by the PM, an initial beta count is taken through a clean section of the filter before any sampling occurs. The reduction in beta count observed after PM has accumulated on the filter is then used to estimate the mass of the particles, with the difference in counts being

proportional to the particulate mass.

A typical beta attenuation system consists primarily of two components: a beta radiation source and a detector. The beta source must meet several criteria—it needs to emit particles energetic enough to pass through the air, the filter medium, and the collected PM; it should provide a strong enough emission rate to ensure reliable counting; it must remain stable over time; and it must pose minimal risk to operators handling the instrument.

The beta gauge functions by recording a baseline count over a clean area of the filter for a fixed duration (commonly around 2 minutes). It then moves that section of the filter to a sampling position, where it collects PM for a set time (usually between 8 to 9 minutes), and finally returns the filter to the detector for a post-sampling measurement. The resulting change in beta count is used to calculate the mass of PM, provided the instrument has been calibrated using a reference filter loaded with a known amount of material. This cycle is typically repeated every 15 minutes to comply with reporting standards.

1.1.9 Datasets

Various datasets on different environmental degradations have been considered to accomplish the goals stated in the thesis. The datasets covered in this thesis include the O-haze and I-haze datasets, covering outdoor and indoor pairs of real-world clean and hazy images used for performing image dehazing. Two Weather, Desnow, and MWI datasets covering Sunny, Cloudy, Snow, Hazy, and Rainy weather images have been used for the multi-weather detection problem. Benchmark datasets such as UIEB, Color-Checker7, SQUID, and SUIM-E are used for performing underwater image restoration. Sample images from such datasets have been shown in Fig. 1.3. Apart from the benchmark datasets, the thesis also covers its own proposed datasets about particulate concentration estimation, details of which are explained in the respective chapters.



O-Haze I-Haze MWI UIEB SQUID Checker7

Figure 1.3: Sample images from multiple benchmark datasets

1.2 The Scope of the Thesis

As mentioned earlier, in this thesis, we have developed methods for performing image dehazing, weather prediction, underwater image restoration, monocular depth estimation, and concentration of particulate matter estimation during day and night using both deep-learning and machine-learning approaches. In the context of underwater image restoration, our methods are solely based on supervised training.

For the dehazing problem in Chapter 2, no training data has been used as the method uses only handcrafted features extracted from the images, followed by the Gaussian-Laplacian pyramid fusion approach to generate the dehazed image, thus working completely in an unsupervised manner. The method initially generates two images based on a white balance and contrast enhancement approach. After two images are derived from the given color image, multiple features are extracted that are highly correlated with the amount of haze in an image. The extracted features are passed through Gaussian and Laplacian filters along with a down-sampling operation, generating multiple layers. This is followed by the fusion of the filtered features with an up-sampling operation to regain the original size of the dehazed image.

Chapter 3 deals with the applications of image dehazing. The first application deals with image-based multi-weather detection as an extension of image dehazing. a novel method to predict different weather conditions that are observed in Kolkata, is proposed in Chapter 3 that takes into consideration “Sunny”, “Cloudy”, “Rainy”, “Hazy”, and “Winter” seasons to be detected from an image. For multi-weather detection systems, the methods rely on a hybrid fusion of handcrafted and machine-based features passed onto a modified support vector machine(SVM) to correctly predict the type of weather from an image, making the process supervised in nature.

The second application deals with estimating the concentration of particulate matter from daytime images. Pollution control boards worldwide give high weightage to the concentration of Particulate matter, specifically of a size less than 10 microns (PM_{10}) and 2.5 microns ($PM_{2.5}$) while computing the Air Quality Index (AQI). The concentration of particulate matter follows a cyclic pattern, exhibiting the impact of seasons on it. We introduce a novel method, detailed in Chapter 3, to estimate the concentration of the particulate matter of the surroundings from an image using machine learning-based methods. This is done in a supervised manner by extracting structural similarity features between an image and its dehazed version, along with the date and time of the snapshot taken. The extracted features are passed to multiple machine-learning methods to make the final estimation, and the one with the highest

accuracy of estimation is selected as the final model.

Chapter 4 deals with estimating the concentration of particulate matter from nighttime images in a supervised manner by designing a CNN-based encoder architecture. Taking the help of a robust dataset covering a wide range of images with different night scenes taken throughout the year, starting from July 2021 to April 2022, a novel deep learning-based model has been explained in Chapter 4 that can estimate the concentration of particulate matter. To the best of our knowledge, the work is the first of its kind, where concentration is estimated for a given region using a nighttime image. The model extracts features such that the spectral reflectance of light caused by the particulate matter and the distance of the objects from the camera to make accurate estimations. To measure a proper depth-map profile, a novel deep learning night image-based depth model has been proposed. Generating an accurate depth map of a color image as well as designing a powerful and complex network to directly regress the depth map is a challenging task. In this thesis, the depth estimation model has been designed to increase the efficiency of the model in order to accurately estimate the concentration of particulate matter using an image of a region as detailed in Chapter 4.

Chapter 5 deals with restoring degraded underwater images with a transformer as its backbone. Restoring the true colors of underwater images due to uneven degradation in the color channels is a big challenge. Floating particles and plankton often introduce noise, complicating many computer vision tasks such as object detection and segmentation. Light intensity diminishes with depth, requiring specialized processing to handle low-light conditions and uneven lighting, which may lead to overexposed or shadowed areas in images. This makes restoring the degraded underwater image a challenging task. In Chapter 5, we propose an 'Aquaformer', transformer-based model designed to process and enhance underwater or aquatic images.

1.3 Conclusions, Limitations, and Future Scopes

This thesis presents a unified vision-based framework for environmental perception under diverse and challenging visual conditions. It begins with the development of an image dehazing system to restore visibility in atmospheric degradation caused by haze and fog. Building upon this enhanced visual clarity, the next chapter introduces a multi-weather detection system that enables robust classification and understanding of dynamic weather conditions. The framework is then extended to the quantitative estimation of particulate concentration from daytime scenes, leveraging visibility cues

and scene depth information. To ensure round-the-clock applicability, the work further explores particulate concentration estimation from nighttime scenes using a novel night-image-based depth estimation strategy. Finally, the thesis advances toward complex underwater environments with an Adaptive Transformer-based approach for underwater image restoration, thereby completing a comprehensive pipeline for visual enhancement and environmental analysis across terrestrial, nocturnal, and underwater domains. We conclude the thesis in Chapter 6. In this chapter, we discuss some limitations of the work presented in this thesis along with possible ways of extending our work in the future.

1.4 The major contributions of the thesis

- An approach for image dehazing using a handcrafted feature-based framework employing pyramidal image fusion using Gaussian and Laplacian filters[25].

  The method initially generates two images based on a white balance and contrast enhancement approach. After two images are derived from the given color image, multiple features are extracted that are highly correlated with the amount of haze in an image. The extracted features are passed through Gaussian and Laplacian filters along with a down-sampling operation, generating multiple layers. This is followed by the fusion of the filtered features with an up-sampling operation to regain the original size of the dehazed image.

- An image based approach for determining different weather conditions *viz.* Sunny, Cloudy, Rainy, Hazy and Winter[26].

  The model extracts handcrafted features that have good correlation with images having a specific season. The sky region in an image has a good influence on the type of weather. To utilize this feature, a deep learning model has been trained to extract sky areas in an image and to further extract features from sky and non-sky regions. The final handcrafted feature vector becomes 93-D after combining features from sky and non-sky regions. The feature vector is passed through a multi-label Support Vector Machine(SVM) to classify the image into one of the designated weather classes.

- Estimation of the concentration of the particulate matter of the surroundings based on image features and date-time information as inputs to machine learning-based methods[27] . The concentration of particulate matter follows a cyclic pattern, exhibiting the impact of seasons on it. Particulate matter specifically of a

size less than 2.5 microns ($PM_{2.5}$) is considered to be among the most vital parameters in gauging the quality of air in the surroundings. Using the effect of the presence of $PM_{2.5}$ on visibility and hence image features supplemented by temporal information, the approach produces estimates of the concentration of the particulate matter using machine learning.

- Estimation of particulate matter concentration from night-time images [28].

  A diverse dataset with wide range of images has been acquired with different night scenes taken throughout the year covering the impact of seasons on the concentration of particulate matters. The model uses features extracted from the multi-scale domain and fuses information extracted from the multi-scale depth map. Depth estimation plays a vital role in many applications such as autonomous driving, 3D scene reconstruction, augmented reality, etc. A novel night-image based depth estimation has been proposed using the features extracted from a CNN-based model complemented by the Conditional Random field (CRF) concept to achieve the objective.The novel model designed to address the challenge uses a fusion of deep learning models along with the concept of Conditional Random Field(CRF), both complementing each other. This helps the models to be specifically used for estimating the concentration during the night scenes.

- Underwater image restoration via Adaptive Transformer[29].

  The major contribution in this work deals with a modified attention mechanism, an adaptive normalization layer, a symmetric padding approach instead of masked cyclic shift, and an adaptive fusion of encoded and decoded features. The four major modifications to the baseline Swin Transformer[30] make the proposed approach suitable for restoring underwater images.

List of Related Publications

Conferences

- **Bhandari, H.**, Palit, S. (2021). Looking Beyond the Haze: A Pyramid Fusion Approach. *Neural Information Processing. ICONIP 2021.*, vol 13110. Springer, Cham. https://doi.org/10.1007/978-3-030-92238-2_54.

- **Bhandari, H.**, Chowdhury, S., Palit, S. (2022). A Hybrid Human-Machine System for Image-Based Multi-weather Detection. *Image and Vision Computing. IVCNZ*

2022, vol 13836. Springer, Cham. https://doi.org/10.1007/978-3-031-25825-1_23.

- **Bhandari, H.**, Palit, S. 2023. REAL-NET: A Monochromatic Depth Estimation Using REgional Attention and Local Feature Mapping. *In Image and Video Technology: 11th Pacific-Rim Symposium, PSIVT 2023, Auckland, New Zealand*, Proceedings. Springer-Verlag, Berlin, Heidelberg, 302–311. https://doi.org/10.1007/978-981-97-0376-0_23.

- **Bhandari, H.**, Palit, S. 2024. "Aquaformer: Underwater Image Enhancement via Adaptive Transformer," 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024.

- Chakraborty, S., Palit, S., **Bhandari, H.**, 2024. "Can an image tell the tale: Looking beyond the haze to determine PM2.5 concentration," 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024.

Journals

- "Dark-AIRNET: A Nighttime Image-based Air Particulate Concentration Estimation Network" accepted by the Journal of Earth System Science.

Chapter 2

Image Dehazing: Pyramid Fusion Approach

2.1 Introduction

Images captured in hazy conditions often suffer from reduced clarity and visibility. This degradation is primarily caused by the scattering of sunlight due to particles and water vapor present in the atmosphere. Computer vision algorithms usually require the input images to be free from degradation, including haze. As a result, effective dehazing techniques are essential for removing haze and restoring image quality. The process of mitigating the effects of haze in foggy or hazy images is referred to as Image Dehazing. An example of this process, utilizing the proposed method, is demonstrated in Fig.2.1. As a result, various dehazing techniques have been developed to recover the original, haze-free image. As haze is not uniformly distributed across an image, and the distance between the camera lens and the different objects constituting the scene varies, this problem cannot be treated as a simple degradation issue. Therefore, an accurate estimation of the distance of the hazy objects from the camera *viz.* the depth or density of the haze, is necessary for effective removal. Several machine learning algorithms have been introduced that leverage prior knowledge, such as non-hazy ground-truth images, to improve the quality of hazy images [31]. Additionally, various filtering techniques have been designed to enhance image contrast [32]. Numerous algorithms have also been proposed to estimate atmospheric models, utilizing reverse techniques to recover clear images from hazy ones [33, 34].

A new approach to image dehazing has been proposed, inspired by various characteristics observed in hazy images. The primary goal is to develop a method that

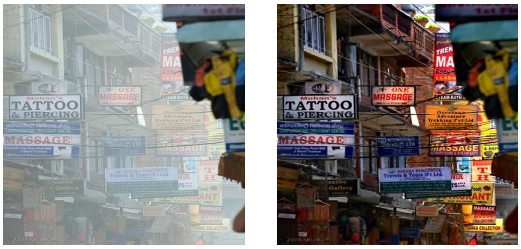
does not rely on any prior knowledge of hazy images or atmospheric parameters. The atmosphere contains particles that alter the passage of light, leading to reduced visibility and making object detection more challenging. Efficiently modeling the process of light scattering due to these particles is a difficult task because it involves several factors, such as the varying size, type, and distribution of the particles, as well as the wavelength of the incident light. In such conditions, the visibility of an object is influenced by two simultaneous processes: first, the light emitted by the object is attenuated due to scattering in various directions and absorption by dust, smoke, and other particles. Second, the airlight, which originates from the Sun, is scattered by the particulate matter in the atmosphere toward the lens along the line of sight. Moreover, both transmission and airlight are affected not only by weather conditions and the angle of sunlight but also by the concentration of particles in the air. The most commonly referenced model for haze creation was introduced by McCartney in [35]. This model states that the hazy image is a linear combination of direct attenuation and airlight.

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{2.1.1}$$

where $I(x)$ and A refer to the received hazy image and the airlight, respectively. The transmission map $t(x)$ of the object depends on its distance from the camera and the scattering coefficient β of the particles [36] such that

$$t(x) = e^{-\beta.d(x)} \tag{2.1.2}$$

where $t \in \varepsilon[0, 1]$. When $t = 1$, light reaches the camera with no loss or attenuation, while $t = 0$ stands for the complete loss of light, indicating the extinction of the object.



(a) Hazy image (b) Dehazed image

Figure 2.1: Image Dehazing using the proposed algorithm

The proposed dehazed method can be divided into three phases:

- Derivation of white-balanced and contrast-enhanced derived images from hazy

images.

- Feature extraction from the derived images.
- Image Fusion

The rest of the chapter is structured as follows: Section 2.2.3 provides an overview of related research in this domain. Section 2.4 outlines the proposed methodology. Finally, the results obtained from the proposed algorithm and a comparison with existing methods are discussed in Section 2.4.2 with noteworthy observations in Section 2.5.

2.2 Related Work

Numerous dehazing algorithms have been proposed over time. Effective dehazing requires an accurate estimation of the haze depth, as the degradation caused by haze tends to increase with the distance from the camera lens to the object. Previous works in this area have often utilized multiple images of the same background. Fusion-based methods are also widely adopted to remove the haze and enhance the contrast of hazy images. The work history has been categorized into three categories.

2.2.1 Traditional approach

Fang [37] introduced a dehazing method that takes advantage of polarization effects, jointly considering the airlight and the object radiance. Gibson and Nguyen [15] provided a metric to estimate the contrast enhancement and human observations to solve the dehazing problem. However, it has a major limitation: it can only assess the images that have been trained to estimate contrast degradations, specifically the foggy ocean scenes. Hence, visibility enhancement without any prior knowledge is of crucial importance. Tan [38] presented a technique to improve the quality of an image degraded by severe weather, based on two key assumptions: first, that non-degraded images exhibit higher contrast, and second, that changes in airlight are generally smooth. He, in [34], developed a dehazing algorithm that relies on the assumption that pixels in non-haze regions have low intensity in at least one color channel. The assumption was called Dark Channel Prior (DCP). In addition, multiple variants of DCP are proposed, such as color-line prior (CLP) by Fattal [39] based on the separation of object shading from medium transmission and haze-line prior (HLP)[40]. Berman in [18] proposed

non-local dehazing approach (DNet) that estimates transmission using global color-line structures in RGB space, enabling robust haze removal in sky and bright regions where Dark Channel Prior typically fails. A more efficient approach involves using only a hazy image for dehazing; however, estimating fog density from a single image remains challenging. Consequently, most dehazing algorithms estimate a transmission map to improve the visibility of hazy images, often utilizing Koschmieder's atmospheric scattering model [41]. Bovik [19] introduced a novel method called the Fog Aware Density Evaluator (FADE), which is based on certain statistical features related to fog. FADE solves the de-hazing problem by adapting enhancement parameters based on real fog density.

2.2.2 Multi-result fusion approach

Image fusion is a technique that combines multiple images into one by integrating their most relevant features, as noted in [42]. A key benefit of using multiscale fusion for image dehazing is its speed, since it doesn't rely on an exact depth map. However, the absence of a reference fog-free image presents challenges in designing such algorithms. Traditional fusion-based methods typically operate on specific input types and tend to perform well in certain areas while showing weaknesses in others. By leveraging the strengths of different processing strategies, one can fuse their outputs to mitigate individual flaws and achieve a more visually pleasing result—this is referred to as multi-result fusion. In [43], Ancuti et al. introduced a multiscale fusion approach for removing haze by carefully selecting input images and designing weight maps that effectively enhance scene visibility, without needing a clean ground truth image. While the FADE technique can significantly reduce haze, it may also lead to unnatural color artifacts in the final image. Li et al. [44] improved fusion by considering factors such as dark channel priors, saliency, and image clarity, and they applied both global and local enhancement techniques. Meanwhile, Ren et al. [45] processed the input image with white-balancing, gamma correction, and contrast adjustments, and then used a multiscale neural network to generate weight maps for fusion. Their model was trained using a mix of MSE loss and adversarial loss to ensure high-quality results.

2.2.3 Multi-feature fusion approach

Beyond image-level fusion, some approaches focus on extracting diverse features from hazy images and applying advanced processing and fusion techniques at the feature

level. Most of these feature fusion strategies rely heavily on convolutional neural networks (CNNs), as CNNs are inherently adept at capturing a hierarchy of features, from low-level patterns like edges and textures to high-level semantic information. This makes them especially well-suited for tasks involving feature-based operations. For instance, Tan et al. [46] designed a network that combines shallow features, derived from image contours, with deeper semantic features obtained through dilated convolutions. These features are then refined and decoded to produce a dehazed image.

In real-world scenarios, a single image often fails to fully capture the scene due to limitations like occlusion or poor visibility. Hence, incorporating data from multiple sources tends to yield a more complete and informative representation. Fusion techniques are developed to gather and reorganize valuable information from various inputs to construct a clearer and higher-quality output. The method proposed here leverages multi-scale feature extraction from different representations of the original hazy image and then fuses them to enhance the final visual result.

2.3 Proposed Methodology

Initially, features that can be computed from intermediary or pre-processed images are selected, with the primary goal of differentiating between natural images and those impacted by haze. A fusion strategy is then employed to combine these features, enhancing the visual quality of the hazy image I.

2.3.1 Formulation of the approach

Two images are generated from the original hazy image. The first image I_{WB} is obtained by applying white balancing to the original image. The second image I_{CE} is created by enhancing the contrast of the original image, as described in Algorithm 1. It is important to note that directly using the CLAHE algorithm or any other histogram equalization technique would introduce artificial color distortions into the output image, and, therefore, such methods are avoided in this approach.

Histogram Equalization (HE) globally redistributes intensity values and increases overall contrast. Treats haze as normal low-contrast, thus enhances haze instead of removing it. CLAHE works locally instead of globally, limiting extreme amplification and producing better results than HE. Although CLAHE improves visibility, it does not restore scene radiance.

Steps 1–3, use luminance as a structural guide normalizes each RGB channel by its mean. This suppresses the white airlight and the color cast and enhances dark objects hidden by haze. Steps 6–7, explicitly re-inject lost spatial details and object boundaries, thus performing real spatial dehazing and not just contrast stretching. Step 8 performs multi-scale information recovery, where the wavelet fusion separates low-frequency illumination and high-frequency edges, selectively combining the best features. Thus, preserves sharp edges, balanced brightness, and noise suppression. Finally, HE is applied to enhance visibility in a safe manner without haze or noise. A sample example with comparison with histogram equalization and CLAHE has been shown in figure 2.2.



a b c d

Figure 2.2: Contrast Enhancement: (a) Original Image, (b) HE, (c) CLAHE (d) Proposed

As previously mentioned, features that are strongly correlated with the level of haze in the image are extracted from each of the processed input images. These images and the corresponding features are then combined through a fusion operation to produce the final dehazed image.

Since haze density is not uniform across the image, applying dehazing to the entire image does not yield optimal results. To address this, a patch-based dehazing approach is proposed. The assumption is that both the depth of the scene and the transmittance remain constant within each patch, provided that the patch size is appropriately chosen. However, dehazing does not significantly affect regions with smooth textures or low contrast as these regions do not contribute meaningfully to the image's overall quality after processing. Therefore, when performing dehazing, regions with

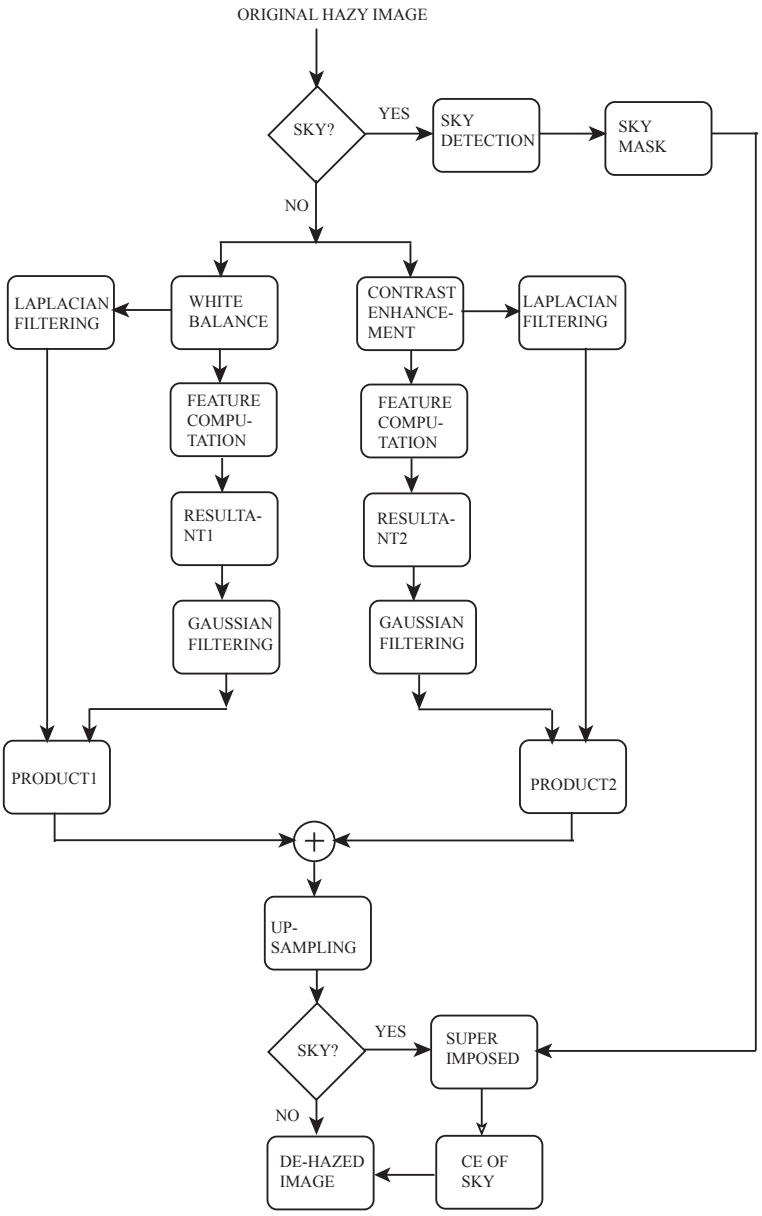


Figure 2.3: The block diagram of the proposed approach

Algorithm 1 Contrast Enhancement

Input: Hazy Image I
Output: Contrast Enhanced I_{CE}

1: $I_{gray} \leftarrow rgb2gray(I)$
2: $MeanIm_c \leftarrow Mean(I_c)$ where c ϵ [r,g,b]
3: $K_c \leftarrow (I_c \times I_{gray})/MeanIm_c$
4: $I_1 \leftarrow concat(K_c)$
5: $I_2 \leftarrow imadjust(K_c)$
6: $G \leftarrow imgaussfilt(I_1)$
7: $J \leftarrow I_1 + (I_1 - G)$
8: $X \leftarrow WaveletFusion(I_2, J)$
9: $I_{CE} \leftarrow histeq(X)$

smooth textures are excluded from the process.

Outdoor hazy images generally consist of both sky and non-sky regions. Sky areas are predominantly smooth, lacking texture and edges, violates the common DCP priors[47]. Aggressive dehazing can introduce artifacts such as color shifts, banding, or unnatural gradients and a slight contrast changes in smooth sky areas are less noticeable than similar changes in textured regions thus no significant improvements is sen when dehazing algorithms are applied on sky. Hence, the dehazing algorithm is applied specifically to the non-sky regions. To separate the sky from the non-sky regions, a mask is created using a sky detection algorithm, such as the one in [36]. In cases where sky regions are minimal, the sky-non-sky segmentation step is skipped, and features are directly extracted from the entire hazy image.

As illustrated in Fig. 2.3, the input hazy image undergoes preprocessing to generate both a white-balanced image and a contrast-enhanced image. The white balancing process removes color inconsistencies, ensuring that objects that appear white in real life are also rendered as white in the image. The contrast-enhanced image is produced using Algorithm 1, rather than using histogram equalization or CLAHE, which are more suited for images with light haze, distributed uniformly all over. Thus, contrast enhancement serves as an additional preprocessing step to improve the final quality of the dehazed image. The feature extraction for the hazy image is performed on non-overlapping patches. It is assumed that the transmission value t of the hazy image varies across the image due to the non-uniform haze density, which also changes with the depth of the objects present in the image.



Figure 2.4: The overall process of the proposed algorithm is illustrated using a hazy image as an example. In the first column, the input hazy image is displayed. The second column contains the preprocessed images, with the top image being White-Balanced and the bottom one showing the Contrast Enhanced version. Columns three through seven present the feature maps for the corresponding derived images, representing Chromaticity, Colorfulness, Entropy, Saliency, and Visibility, respectively. The eighth column shows the normalized feature map generated from each input-derived image. Finally, the dehazed output is obtained by applying the fusion process to the two resultant images, as shown in Fig. 2.1(b).

2.3.2 Feature Extraction

As illustrated in Fig. 2.3, for an image with minimal sky region, two images are generated: the White Balanced Image and the Contrast-Enhanced Image. These images are collectively referred to as input-derived images, with k=1 denoting the White Balanced image and k=2 corresponding to the Contrast Enhanced image. Five window-based attributes — Colorfulness, Visual Saliency, Entropy, Visibility, and Chromaticity are computed for each of the derived images. The spatial distribution of each attribute across the input-derived images forms the corresponding feature map for that attribute. Normalized weight maps for each attribute are obtained by scaling the weight maps of both input-derived images to match the scale of the original input image. The computation of the features and their corresponding weight map is now explained.

Colorfulness

Haze in an image results in a degradation of its color and alters the scene's overall hue, leading to poor visual perception. Colorfulness is defined as the degree of difference between the gray and colored aspects of the image. It is expressed as a linear combination of the local mean and standard deviation. These parameters are computed as described in Eq. (2.3.1) by dividing the image into small patches. Fig. 2.5 presents the results of this analysis by considering non-overlapping 5×5 patches for both the clear

image and its corresponding hazy version. Fig. 2.5(e) demonstrates the relationship between the sample standard deviation and the sample mean for an image, as shown in Fig. 2.5(a). The presence of fog results in a shift in the scene's color, as indicated by a decrease in both the local mean and local standard deviation compared to the ground truth, which reduces colorfulness.

$$ColorFulness^k = w_1 \cdot \sqrt{\sigma_a^2 + \sigma_b^2} + w_2 \cdot \sqrt{\mu_a^2 + \mu_b^2} \tag{2.3.1}$$

where σ_a and σ_b represent the standard deviations, and μ_a and μ_b are the sample means of the channels a and b of CIEL*a*b colorspace, respectively while w_1 and w_2 are the weights assigned to the sample standard deviation and sample mean, respectively.

Visual Saliency

The human visual system assigns varying levels of attention to different objects and areas in a scene, with regions containing more distinct and well-defined objects attracting greater focus. Haze in the scene tends to make closer objects more noticeable, as saliency decreases sharply with even a slight increase in depth. The saliency of an image emphasizes the objects in the scene relative to their background. Visual saliency is computed by combining three priors derived from the color image: color, frequency, and location.[48].

$$\begin{aligned} Saliency^k &= SF.SC.SD \\ SF &= (I_L * g)^2 + (I_a * g)^2 + (I_b * g)^2 \\ SC &= 1 - \exp(-\frac{(I_a)^2 + (I_b)^2}{\sigma_c^2}) \\ SD &= \exp\left(-\frac{(k-c)^2}{\sigma_d^2}\right) \end{aligned}$$

where I_L, I_a and I_b represent the three channels derived from the color image I in the CIE L*a*b* colorspace. σ_c and σ_d are parameters, c denotes the center of the input image window W_k with normalized values of I_a and I_b being used in the equations. g refers to a 2-D circular symmetric log-Gabor filter. Additionally, objects closer to the central region of the image are more prominent in human perception, making them more likely to be salient compared to those farther from the center. The construction of the Visual Saliency Map involves calculating the Visual Saliency as outlined by the

following equation (2.3.2).

$$VisualSaliency^k = Saliency^k * G_{r,\sigma} \tag{2.3.2}$$

where $G_{r,\sigma}$ is a Gaussian Filter with mean r and sigma σ.
As shown in Fig: 2.5(d), a histogram of the saliency map for the images in Fig: 2.5(b) which feature natural scenes with varying levels of haze, reveals that as haze intensity decreases, the saliency distribution tends to follow a more regular pattern with a reduction in skewness. In contrast, the histogram of the saliency map for hazy images displays a more irregular pattern, with an increase in distortions as haze density rises. This indicates that the saliency map effectively captures the distortions introduced by the presence of haze in the image.

Entropy

The entropy of an image measures the amount of information it contains. As the density of fog increases, the entropy of the image decreases. The entropy feature map captures variations in the gray-level distribution of the image. The image is divided into multiple patches, with each patch providing distinct information. With $M \times N$ representing the window size, the entropy of a given window can be calculated in the spatial domain. This window-based entropy calculation applied across the entire image results in its corresponding weight map.

$$Entropy^k = \sum_{i=0}^{L-1} p_i \log \frac{1}{p_i} \tag{2.3.3}$$

where p_i = probability distribution of ith intensity level in the window with L gray levels.

Visibility

The visibility of an image refers to its clarity. A color image can be either transparent or non-transparent, depending on the medium. Visibility is associated with the maximum distance at which an object of a certain size can be observed and recognized by the observer against the sky horizon. The visibility feature map is determined using a blur estimation method. Initially, the color image is blurred using a Gaussian function, as shown in Eq. (2.3.4). Then, the root squared blurred difference between the color

image and its blurred version is computed to assess the visibility, as detailed in (2.3.5).

$$\begin{aligned} I_{blur} &= I(x,y) \otimes G_M(x,y,\sigma_1) && (2.3.4)\\ V_1 &= \sqrt{(I(x,y) - I_{blur}(x,y))^2 \otimes G_M(x,y,\sigma_2)} \\ V_2 &= rgb2gray(V_1) \\ Visibility^k &= \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} \frac{V_2(i,j) - \mu}{\mu^{\alpha+1}} && (2.3.5) \end{aligned}$$

Here $G_p(x,y,\sigma)$ represents a circular symmetric weighting function with a patch size of $p \times p$ and a standard deviation of σ. The operator $\otimes$ denotes a 2D convolution operation. $V_2(i,j)$ refers to gray intensity at position (i,j). α is a visual coefficient and μ is the average gray intensity within the patch.

Chromaticity

Chromaticity refers to the quality of color, which includes both hue and saturation. While fog in an image does not significantly affect the hue, it greatly reduces the saturation. As the density of haze increases, the saturation of colors diminishes, resulting in a reduction in chromaticity. To assess this decrease in saturation, the chromaticity feature map, calculated using a window W of size $M \times N$ as described in (2.3.6).

$$Chromatic^k = \exp\left(-\frac{(S_j - S_{max})^2}{2\sigma_e^2}\right) \quad (2.3.6)$$

S_j represents the saturation of the jth patch, S_{max} is the maximum saturation of the corresponding patch, and σ_e denotes the standard deviation.

2.3.3 Image Fusion

Weightmaps corresponding to the different features for each of the derived images from the input image, *viz.* the White-balanced and Contrast-enhanced ones, are combined to produce a single, enhanced image.

The normalized value of the jth feature($j \in (1,5)$) for each patch i is calculated as shown in equation (2.3.7)

$$W_{k,j}^i = \frac{feature_k^i}{feature_1^i + feature_2^i} \quad (2.3.7)$$

To ensure consistent results, each feature map is derived by normalizing the fea-

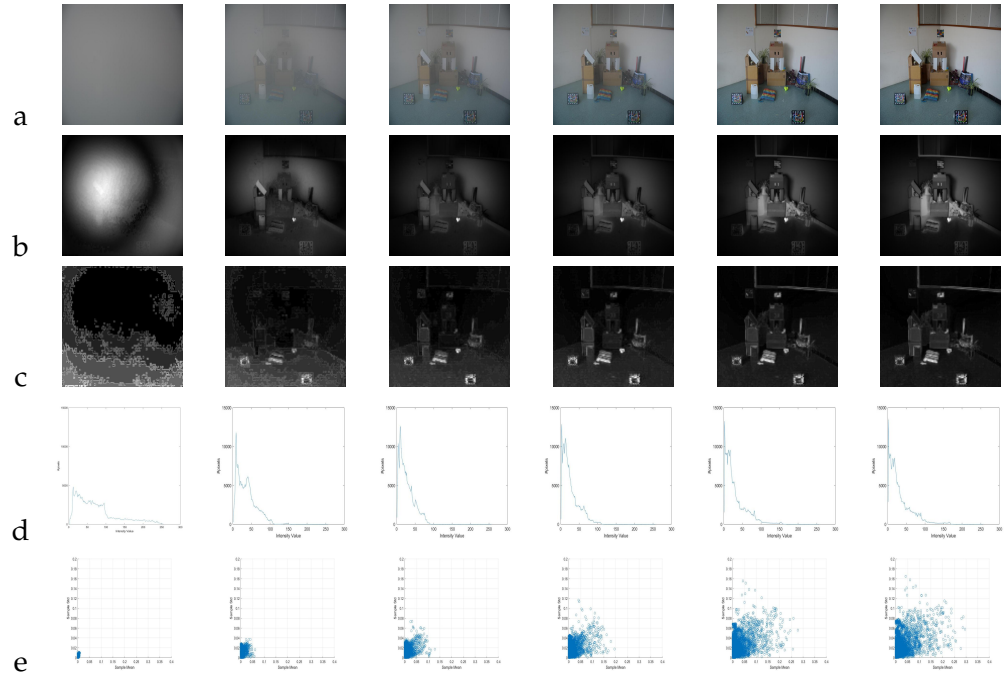


Figure 2.5: Variation in the Saliency and Colorfulness Map using the CHIC Database [1] with the haze level increasing from right to left, where the last column represents the ground truth

tures computed for each input. As a result, two feature maps are obtained, one for each of the derived input images, by combining the five features as described above. The resultant feature map is shown below:

$$Resultant feature Map_k = \prod_{i=1}^{5} \lambda^i W_k^i \tag{2.3.8}$$

where λ^i refers to a multiplicative factor given to each weight map.
Directly applying the normalized weight maps can introduce halo artifacts. To mitigate this issue, a multi-scale Gaussian bandpass filter with downsampling is applied to each of the derived input images I_k, while a bandpass down-sampling Laplacian filter is applied to the resultant weight map W_k to generate fusion output with l levels as shown in (2.3.9). The fusion is then performed in a pyramid-hierarchical fashion. The sizes of the frequency bands decrease as the decomposition level increases. At each level, the Laplacian weights and Gaussian-derived inputs are fused and then up-sampled separately to generate an inverse fused Laplacian pyramid, as shown in Fig. 2.6.

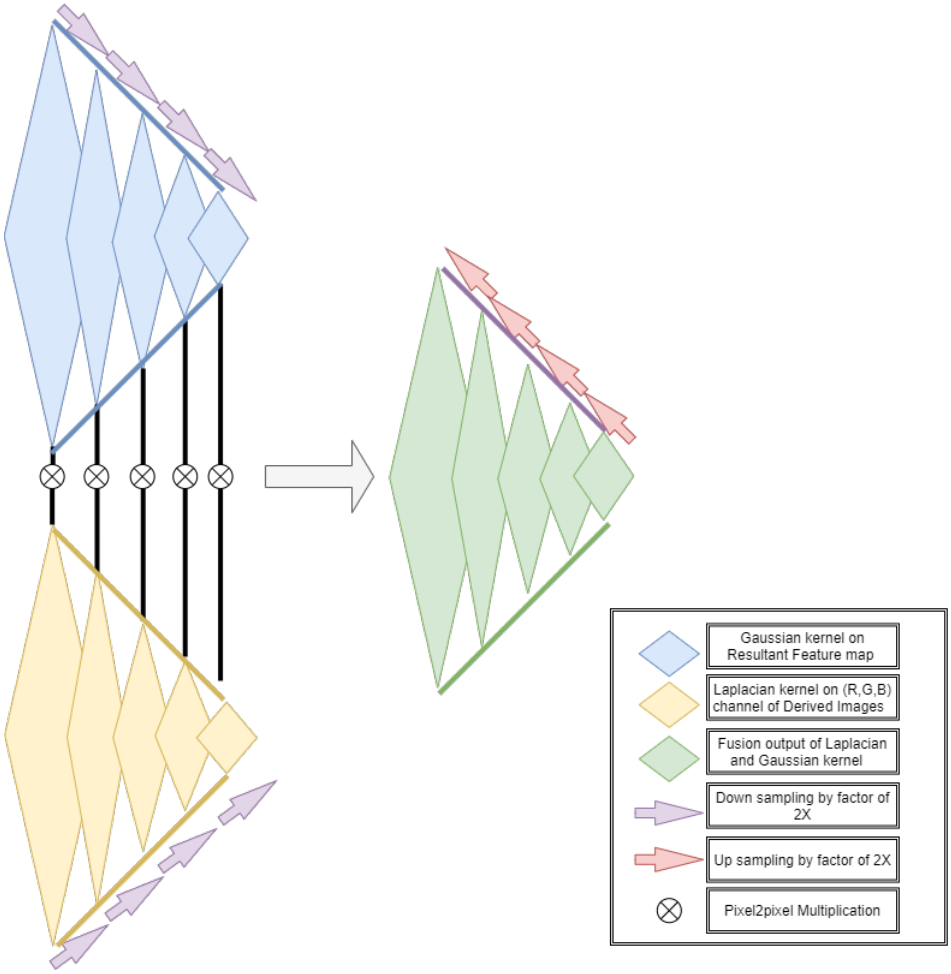


Figure 2.6: The block diagram of the Pyramid Fusion approach

$$F_l = \sum_k \{G_l\{W^k\}\} L_l\{I_k\} \tag{2.3.9}$$

Here, l represents the pyramid levels. The value of $l = 7$ was chosen experimentally to generate the Gaussian and Laplacian levels, denoted as G_l and L_l respectively, with operations applied successively on each level in a bottom-up manner. Ultimately, a dehazed image is obtained by performing Laplacian pyramid fusion using an upsampling operator with a factor of 2^{l-1}.

2.4 Experiments

In this section, we present the experimental setup, results, and datasets. The performance of the proposed dehazing method is evaluated through both qualitative and quantitative analyses. For this evaluation, several benchmark datasets are utilized, and the results are compared with those of state-of-the-art methods, whose codes are publicly available. The quantitative evaluation of the dehazing technique is performed using three metrics outlined in [49]. Metric $\mathbf{e} = \frac{n_r - n_o}{n_o}$, computes the ratio of new edges visible in the dehazed image after enhancement. Here, n_r = number of visible edges in the restored image and n_o = number of visible edges in the original foggy image. So, e>0 indicates improved visibility, e = 0 means no new visibility, and e<0 refers to lost visibility. Metric $\sum = \frac{n_s}{width \times height}$ represents the percentage of pixels that become black or white after dehazing. Here, n_s = number of pixels that became black or white after enhancement, and width and height being the image width and height respectively. It measures the information loss due to over-enhancement and detects washed out regions, over-saturated sky, and burnt highlights. $\sum \approx 0$ means safe enhancement, $\sum > 1 - 2\%$ denotes risk of visual degradation, and large $\sum$ indicates severe loss of information. Metric $\bar{r}$ measures the mean ratio of the gradient before and after dehazing. It measures how much stronger each visible edge became. $\bar{r} \approx 1$ means no improvement with $\bar{r} > 1.5$ leading to good enhancement, and $\bar{r} > 2$ indicates strong visibility restoration. Additionally, patches of varying sizes were used to assess the performance of the algorithm. Peak-Signal-to-Noise ratio (PSNR) and Structural Similarity (SSIM) are metrics used mainly for images with clear reference images, but the real world haze problem lacks such reference data, and hence the three metrics are used to qualitatively analyse the method.

2.4.1 Datasets, Parameter Settings & Simulation Environment

The dehazing model has been tested on two benchmark datasets, namely O-Haze[2] and I-haze[3]. Based on the feature weight maps, the parameters used in our approach were empirically tuned as $M = 11$, $N = 11$, $\omega_1 = 0.3$, $\omega_2 = 1$, $\sigma_c = 0.25$, $\sigma_d = 114$, $\alpha = 0.65$, $\sigma_e = 0.7$. The Matlab program is simulated under the stated computer configuration: Intel(R) Core(TM) $i7 - 9700KU$ CPU @ 3.6GHz, 8 Core(s) with $32GB$ RAM, having version MATLAB 2019b.



Figure 2.7: A comparison of single image dehazing algorithms for heavily hazed images is conducted using the O-haze Database[2] and the I-haze Database[3]

Foggy Image	**AMEF[15]**			**CAP[16]**			**BCCR[17]**		
	e	$\sum$	$\bar{r}$	**e**	$\sum$	$\bar{r}$	**e**	$\sum$	$\bar{r}$
a	20.4962	0%	2.0984	15.6188	0%	1.495	37.8797	0%	**7.1776**
b	0.33001	0.01%	1.8823	0.4817	0.0017%	0.88576	**1.4667**	0.00027%	2.3822
c	5.8359	0%	2.1207	4.0596	0%	1.5607	12.6996	0.0016%	3.5014
d	1.192	0%	2.0108	1.0043	0%	1.294	**3.8736**	0%	4.1512
e	3.2154	0%	2.1277	2.5562	0%	1.6384	6.3986	0%	4.0802
f	3.2429	0%	1.9643	2.5562	0%	1.6403	9.5827	0%	**4.977**
g	6.6508	0%	2.3786	4.3939	0%	1.7385	11.7851	0%	3.8604
h	0.56542	0%	1.1719	0.49842	0%	1.5174	1.2998	0%	**5.2241**

Table 2.1: Comparison of Dehazed Images with AMEF[15], CAP[16], and, BCCR[17] shown in Fig. 2.7

Foggy Image	**DNet[18]**			**Defade[19]**			**Proposed**		
	e	$\sum$	$\bar{r}$	**e**	$\sum$	$\bar{r}$	**e**	$\sum$	$\bar{r}$
a	6.2432	0%	1.4112	0.6705	0%	2.2222	**43.1058**	0%	4.5276
b	0.19966	0.088%	1.1218	0.69269	6.9322%	1.3489	1.1477	0.02%	**2.4341**
c	2.5536	0%	1.5169	9.9015	0%	2.3326	**14.76**	0.003%	**4.7548**
d	0.45841	0%	1.3613	2.1879	0.285%	1.7994	3.2414	0%	**7.537**
e	1.5127	0%	1.5072	0.81787	0%	1.7042	**7.0074**	0%	**7.1746**
f	1.5318	0%	1.4658	3.3081	0%	1.9235	**17.6243**	0%	3.0381
g	2.8093	0%	1.6081	9.1045	0%	1.8996	**11.9848**	0%	**4.512**
h	0.29086	0%	1.3336	0.6705	0%	2.2222	**1.3982**	0.0006%	3.7427

Table 2.2: Comparison of Dehazed Images with DNet[18] and Defade[19] models shown in Fig. 2.7

2.4.2 Performance Analysis

As can be observed from Table 2.1 and 2.2, taking into consideration severely dense, hazy images, the proposed algorithm restores the image in a much better way than many of the state-of-the-art algorithms. Results obtained from [17], though showing good results, face a major drawback, i.e., the selection of the haziest region in the image has to be done by the user, thus making it an interactive algorithm.

2.4.3 Limitations

The proposed dehazing method works well in image with homogenous haze but fails in scenarios where the presence of haze in images is heterogenous in nature i.e. a region in an image contains haze and other region is hazefree. Further, the model fails to produce clear hazefree output for images with extreme dense haze as the Gaussian and Laplacian filtering is in sufficient to tackle such noise.

2.5 Concluding remarks

A new and unique framework for the dehazing problem has been presented. An important limitation of state-of-the-art algorithms was observed to be the presence of artificial colours in the de-hazed image especially for heavy and non-uniform haze. This has been effectively addressed through a number of innovative strategies. Deriving the white balanced and contrast enhanced images served to decouple the issues of colour preservation and haze removal. Identification of sky regions as having little significance towards the computation of haze features aided in reducing computation. Finally, applying pyramid fusion to combine the various feature weightmaps was a crucial step in preventing undesired artifacts in the final de-hazed image. The experimental results indicate that the developed algorithm produces good results as compared to many state-of-the-art algorithms, even though no complex networks requiring training images in large numbers have been deployed.

Use of machine learning is known to produce improvements over the use of handcrafted features when a large amount of data is available. The next chapter deploys machine learning strategies to develop two important applications based on de-hazing.

Chapter 3

Application of Machine Learning using image-based features for Multi-Weather Detection and Particulate Matter Concentration Estimation

3.1 Introduction

A set of useful image-based features were developed in Section 2.3.2 of the previous chapter. These are now used in conjunction with machine learning to develop solutions to two important applications – first, determining the weather and second, computing the daytime concentration of atmospheric particulate matter, both relying on the input image as the only source of information.

The first application derives its importance from the fact that utilizing images for accurately predicting weather conditions is crucial for meteorological forecasting and various other uses, including the driving of autonomous vehicles, air quality monitoring, and more. Determining the environmental conditions is vital for the safe operation of autonomous vehicles, which can be severely affected by adverse weather such as haze, rain, snow, or strong sunlight. The strategy to be employed for enhancing them depends heavily on the type of degradation in the image. Since autonomous vehicles may encounter different weather conditions, it is essential to accurately and quickly determine the weather from the camera output.

The second application, estimating the concentration of particulate matter, particularly fine particles like $PM_{2.5}$, from images of the surrounding environment, is especially important, considering the growing threat of air pollution. Further, rapid industrialization and urbanization in recent years have significantly increased concerns about toxic emissions into the atmosphere. Air pollutants, particularly fine particulate matter ($PM_{2.5}$) and suspended particulate matter (PM_{10}), nitrogen oxides, and sulfur oxides, pose serious environmental and public health risks. Among these pollutants, $PM_{2.5}$ is considered the most hazardous, especially concerning human health. Inhalation of elevated $PM_{2.5}$ levels is linked to respiratory diseases, physiological dysfunctions, and irritation of nasal and ocular tissues, highlighting the urgent need for effective monitoring and control of these pollutants.

Traditional methods of monitoring air quality, such as Air Quality Monitoring Stations, are effective but come with several drawbacks. They are expensive, require qualified professionals for operation and maintenance, and are not easily scalable. As a result, there is a growing need for cost-effective, portable alternatives for monitoring $PM_{2.5}$ concentrations, particularly in urban and industrial areas.

The image-based multi-weather detection system is discussed in Section 3.2. Section 3.2.1 reviews the concerned related work, while Section 3.2.2 discusses the human visual features emulated in the image-based weather detection system. Section 3.2.3 discusses the Convolutional Neural Network used for the task at hand. Section 3.2.4 presents the various datasets used in the study and a detailed explanation of MLJSVM. Finally, Section 3.2.5 deals with the ablation studies and comparison of the output produced by the proposed and other state-of-the-art methods for weather detection.

The daytime image-based particulate estimation system is explained in Section 3.3. A review of work in the area is described in Section 3.3.1. Section 3.3.2 presents the details of the proposed approach. Section 3.3.3 presents the novel data set designed and used in the study and comparison of the output produced by different machine learning models for the estimation of particulate concentration. Finally, notable observations are stated in Section 3.3.4.

In Chapter 2, we focus on using handcrafted features to solve the image dehazing problem. In contrast, Chapter 3 employs handcrafted features with a different objective—weather classification and the dehazing method for particulate concentration estimation. The weather detection system uses the handcrafted features, combined with ResNet-101 generated features and passed to a Multi-class Latent Joint Support Vector Machine for weather classification. By applying a dehazing algorithm, the estimation method utilizes the Structural Similarity Index Measure (SSIM), providing a

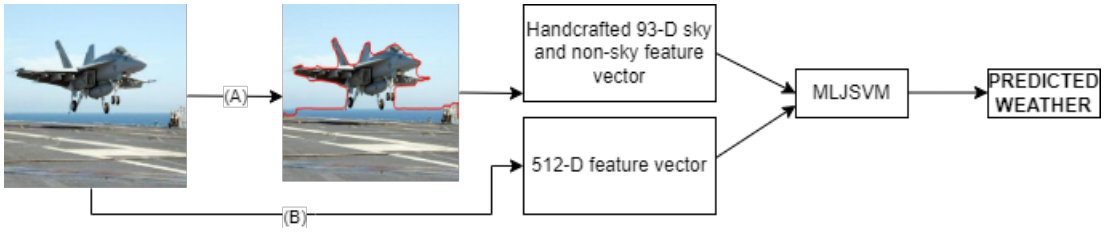


Figure 3.1: The proposed weather detection method includes two main components: (A) the SKYNET model, which generates sky and non-sky segments from an input image, and (B) the advanced ResNet-101 model, which extracts machine-based features. Handcrafted features are extracted from the segmented sky and non-sky regions and combined with the ResNet features. The combined features are fed into the MLJSVM classifier to generate the final output.

more reliable feature for $PM_{2.5}$ estimation.

3.2 Image based Multi-weather detection

In this section, we elaborate on a novel technical insight into multi-weather classification tasks. Our approach draws inspiration from the human visual system, leveraging features that the brain associates with different weather conditions. Provided with an outdoor image, the system can classify the weather as “Sunny”, “Cloudy”, “Rainy”, “Hazy”, or “Snowy”, as shown in Fig. 3.2.

3.2.1 Related Work

The content of this section deals with a discussion of related works, categorized into two main groups based on their area of application. The first group includes applications to extract or compute weather-related information. The second group focuses on scenarios where weather classification and understanding are explicitly utilized for further tasks.

Applications requiring weather information

Despite the vast potential of weather detection from single images, there has been limited research in this area. Most studies have focused on weather classification for a small set of static scenes, and with varying background scenes, such classifiers may fail when applied to the images. [50, 51] have proposed weather recognition from vehicle sensor data and operate within specific regions of interest. To effectively categorize a

scene based on weather, it is necessary to extract structural information, which could involve using illumination-invariant features like SIFT or HOG [52, 53].

Weather information extraction has proven valuable in a wide range of applications. In [54], a convolutional neural network (CNN) was employed to estimate various attributes and features, including the properties of the scene, season, and weather. Glasner [55] examined the relationship between ambient temperature and the visual characteristics of outdoor scenes, analyzing the correlation between measured temperature and image sequences of such scenes. In [56], the time and location of a scene were estimated using information such as sun position and weather conditions, along with scene structure. Additionally, Workman [57] utilized cloud-related information to estimate the geometry of outdoor scenes.

Weather Classification

Using Image-Based Features

Understanding weather conditions is vital for a range of real-world applications, particularly in the context of autonomous vehicles, where safety can be greatly enhanced by adjusting vehicle speed in response to adverse weather situations [50, 51]. Kurihata in [58] developed a method for detecting rainy weather using an in-vehicle camera that identifies raindrops on the windshield. This technique relies on a learned discriminative framework to classify the weather conditions. Martin [50] improved upon this approach by distinguishing multiple weather conditions using monocular color images, though this method was specifically designed for driver assistance and faced limitations in scope. Yan in [51] advanced weather recognition for vehicles using features extracted from images and a Real Adaboost classifier, but similar to Martin's work, the method is limited to specific target scenes and cannot handle a wide variety of weather conditions.

Li [59] focused on distinguishing Sunny and Cloudy conditions in popular tourist spots, but with the approach requiring multiple images of the same scene, limiting its application to a small set of locations. Lu [60] proposed a method for labeling images as either Sunny or Cloudy, and Chen [61] expanded on this to include Overcast conditions as another class. Song [62] proposed a multi-class weather classification technique, but like others, it was designed for a fixed scene. In [63], a Convolutional Neural Network (CNN) was used for classification in conjunction with a Support Vector Machine (SVM).

Narasimhan [64] introduced a model that captures the scattering of light in weather conditions such as fog, haze, mist, and rain, particularly useful in nighttime scenar-

ios due to strong scattering effects. Laffront [65] examined 40 features extracted from an image and developed a model to predict them, while Zhang [66] used multiple weather features combined with multi-kernel learning to create the classification system.

While the methods described above show promising results in their respective applications, they are often constrained by specific conditions or assumptions, which can be significant drawbacks. These limitations highlight the need for more generalized and robust weather detection approaches.

Using CNN

Deep Neural Networks, particularly Convolutional Neural Networks (CNNs), have transformed image classification tasks across various domains, including image and video processing, object detection, and more. CNNs excel at extracting both local and global scene information, which makes them suitable for a wide array of image classification problems. However, weather classification is a unique challenge compared to general object classification, as it requires the extraction of weather-sensitive features that may not always be immediately apparent in the input image.

In the realm of weather classification, CNNs can sometimes struggle to capture subtle yet crucial weather-specific information. For instance, Elhoseiny [67] utilized VGGNet for weather classification, and Lu [60] employed image-based and CNN-extracted features to classify weather images. While these models show promising results, they are inherently limited by their reliance on CNNs, which may overlook key weather-related features embedded in the input image. Moreover, many existing approaches focus on only a few weather conditions, such as “Sunny” and “Cloudy,” limiting their generalizability and effectiveness in diverse real-world scenarios.

Despite these limitations, image-based weather observation presents immense potential for practical applications, especially with the proliferation of low-cost smartphones and surveillance cameras. The availability of accessible imaging technology motivates the search for more intelligent and efficient solutions to weather detection and classification, which can leverage both traditional image processing methods and deep learning techniques. By incorporating weather-sensitive features alongside robust CNN-based approaches, we can enhance the accuracy and versatility of weather classification systems, ultimately paving the way for more reliable and scalable applications in various fields, including autonomous driving, environmental monitoring, and more.

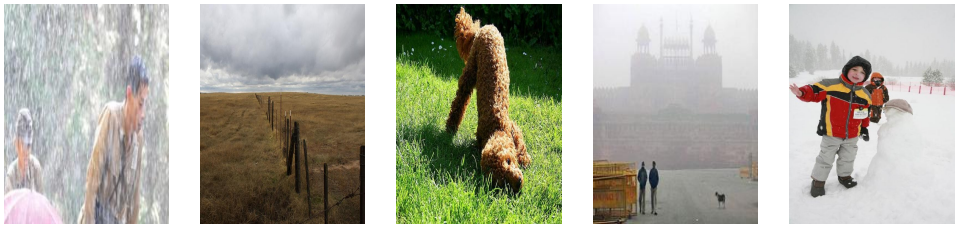


Figure 3.2: The Five weather classes - Rainy, Cloudy, Sunny, Haze, and Snow are listed from left to right, respectively.

3.2.2 Weather Features

Each input image is partitioned into two mutually exclusive segments: sky and non-sky. The non-sky segment undergoes over-segmentation using the Mean Shift clustering [68]. This results in multiple distinct segments within the non-sky region. Features specific to sky regions are extracted from the sky segment.

For each segment in the non-sky area, we compute four weather-related features. These features, along with the sky features, form the overall weather feature vector. Specifically, the feature vector consists of two parts:

1. Sky Feature (f_{sky}): This is derived from the sky segment.
2. Non-sky Features: These include five features, of which
   (a) Visibility (f_{vis}) is computed on a patch-wise basis (using small patches within the non-sky regions).
   (b) The remaining four features $f_{il}, f_{Col}, f_{ce}, and f_{sat}$, are computed directly from the entire image.

The final weather feature vector is obtained by concatenating these six components:

$$[f_{il}, f_{vis}, f_{Col}, f_{ce}, f_{sat}, f_{sky}] \tag{3.2.1}$$

where the features $f_{il}, f_{vis}, f_{Col}, f_{ce}, f_{sat}$, and f_{sky} namely represent illumination, visibility, colorfulness, contrast energy, saturation and sky correspond to vital weather information extracted from the images that will now be described.

Illumination

The proposed method for detecting sunny weather shifts the focus from traditional shadow detection to analyzing how sunlight influences the colors of a scene, partic-

ularly making objects appear brighter and more vibrant. This approach begins by over-segmenting the image using Mean Shift clustering, which divides the image into smaller, more manageable segments, focusing on regions that may exhibit sunlight effects.

A pool with the twenty largest non-sky segments from a collection of sunny images is created. This pool, denoted P_S, contains segments that typically reflect the vibrancy and brightness associated with sunlight. The sky region is excluded from this pool as it doesn't contribute to the sunlight detection process. Additionally, to avoid misclassification of very light segments, such as those that are nearly white, these are filtered out using the L component of the CIE Lab color space of the selected segments. which identifies perceptual lightness.

For each segment in an unknown image, the algorithm calculates the average distance to the K-nearest neighbors (with $K = 5$) from the P_S pool. This calculation, performed in the perceptual color space, helps identify segments that share color characteristics with known sunny segments. A smaller distance between a segment and its nearest neighbors suggests that it is likely part of a sunny scene. The distances calculated form a 20-dimensional illumination feature vector f_{il}, which captures the sunlight-induced brightness and color vibrancy of the segment.

This approach avoids the common challenges of shadow detection, such as misclassifying dark or thin shadows, and ensures that only segments that truly reflect the effect of sunlight are considered. By focusing on perceptual lightness, the method also filters out extreme color segments, like black or white, which might otherwise be incorrectly classified as sunny. As a result, the proposed technique offers a robust and effective solution for accurately detecting sunny weather from images.

Colorfulness

Colorfulness refers to the visual sensation of an image, capturing the degree of chromaticity as sensed by the human eye. It quantifies the difference between a color and its grey counterpart, indicating how vivid or saturated the colors in an image appear. The colorfulness of a segment in an image is influenced not only by its spectral reflectance but also by the strength of illumination. As the illumination increases, the colorfulness of the segment also tends to increase, unless the brightness becomes excessively high.

In the proposed SKYNET model, the color image is segmented into two distinct regions: the sky and non-sky areas, denoted by S and NS, respectively. To calculate the colorfulness of an image, the algorithm follows a specific computation process (similar



Figure 3.3: Highlighting the impact of illumination in detecting sunlight-bearing segments.

to equation (3.2.2), which considers the illumination effects and spectral reflectance of both regions.

$$ColorFulness = \lambda_1 \cdot \sqrt{\sigma_a^2 + \sigma_b^2} + \lambda_2 \cdot \sqrt{\mu_a^2 + \mu_b^2} \tag{3.2.2}$$

where σ_a and σ_b refer to standard deviations and μ_a and μ_b are sample means of the channels a and b of CIE L*a*b colorspace respectively, λ_1 and λ_2 are weights assigned to the sample standard deviation component and sample mean component respectively. For each of the largest 20 segments of the non-sky area of image I, we define

$$\begin{aligned} f_{col}(i) &= w_i * ColorFulness_i(I) \\ w_i &= N_i / N_I \end{aligned} \tag{3.2.3}$$

where w_i is the weight assigned to $ColorFulness_i$ of the segment i of the image I, N_i is the number of pixels of segment i and N_I is the total number of pixels of . This generates a 20-D f_{col} Colorfulness feature vector.

Outdoor images captured under various weather conditions exhibit different levels of global and local colorfulness. The colorfulness of each segment is calculated, and the normalized colorfulness of the twenty largest non-sky segments, based on their size, is determined. While identifying the largest twenty non-sky segments, any segments that are black or white are excluded, as they represent extreme color values. This process results in the creation of a 20-D Colorfulness feature vector, denoted as f_{Col}. For the sky region S, colorfulness is computed without any mean shift, as the entire sky area is considered for evaluation.

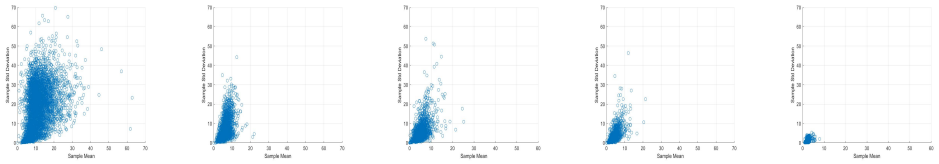


Figure 3.4: Colorfulness of images (a) Sunny (b) Cloudy (c) Rainy (d) Snow (e) Hazy

Segments from sunny images generally display the highest colorfulness due to the vibrant effect of sunlight, which enhances the colors. On the other hand, snow and hazy images tend to have the lowest colorfulness because visibility is significantly reduced in these conditions. In rainy and cloudy images, the scene's color shifts, leading to a reduction in colorfulness. This trend is depicted in Fig. 3.4, which illustrates the colorfulness impact across the images in the weather datasets mentioned in Section 3.2.4.

Presence of Haze

Extensive research has been conducted on haze in computer vision, with the dark channel prior (DCP) being widely recognized as a useful feature for haze detection. However, real-world images often show that the DCP feature does not fully capture the impact of haze. This limitation stems from the assumption in DCP that at least one of the color channels in [R, G, B] is zero or close to zero in the original, non-hazy image. Since haze is typically white, it raises the values of all color channels, making this assumption less accurate in real-world scenarios. In practice, haze often appears with a bluish tinge, as seen in Fig. 3.5(d-f), where the DCP values are significantly lower compared to the images in Fig. 3.5(a-c), which more accurately represent white haze. Additionally, reddish or yellowish haze tints have also been observed. Another limitation of the DCP approach is its inability to differentiate between images with varying levels of haze, as shown by the similar DCP values in images (a) and (b), which have different haze intensities.

The most important factor in a hazy image is the scene visibility. Haze typically forms when dust and smoke particles accumulate in dry air, particularly during winter or colder temperatures, and settle at lower altitudes, reducing visibility. To address this, we estimate the visibility of an image by calculating its blur, as described in equation (3.2.4), and use this estimation to develop a visibility measure, as shown in equation (3.2.5). It may be noted that this is the same Visibility measure of equation (2.3.5).



Figure 3.5: The DCP maps are displayed in the second row, with the first three columns showing images with white haze and the last three columns showing images with non-white haze.

$$
\begin{aligned}
I_{blur} &= I(x,y) \otimes G_p(x,y,\sigma_1) && (3.2.4)\\
V_1 &= \sqrt{(I(x,y) - I_{blur}(x,y))^2} \\
V_2 &= rgb2gray(V_1) \\
Visibility &= \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{V_2(i,j) - \mu}{\mu^{\alpha+1}} && (3.2.5)
\end{aligned}
$$

Here $G_p(x,y,\sigma)$ represents a 2D Gaussian circular symmetric weighting function with a patch size of $p \times p$ and standard deviation σ. The symbol $\otimes$ denotes a 2D convolution operation. $V_2(i,j)$ refers to the gray value at the pixel position (i,j), while α is a visual coefficient, set to 0.65, and μ is the mean of the gray values within the window. Most haze regions exhibit smoothness, which increases as the haze density grows. The image is divided uniformly into overlapping regions, with patch sizes N set to 5, 7, 9, 11, 13, 15, and 17. The median of the median values of the visibility map across these patch sizes is then computed to form a 7-dimensional f_{vis} feature vector.

The Visibility map for both hazy and haze-free images is shown in Fig. 3.6. The measure is observed to perform effectively, even in the case of white haze.

Sky Segmentation

The presence of the sky in an image provides crucial information for determining weather conditions. However, the texture and color of the sky are influenced not only by the prevailing weather but also by the time of day. Sky segmentation is a challenging task because the shape and size of the sky area can vary significantly be-



Figure 3.6: Scenes with clear and natural haze are presented, where the top row contains both hazy and haze-free images, and the bottom row shows their corresponding Visibility Maps.

tween images, and the region of interest may be discontinuous or overlap with similar non-sky regions such as buildings or other structures.

Unfortunately, no existing algorithm can consistently and robustly segment the sky from images under diverse weather conditions, nor is there a publicly available dataset that includes images captured in various weather types along with sky masks as ground truth. Some previous methods, like those proposed in [60, 69], assume only two weather conditions for segmentation and rely on traditional methods. These conventional techniques using handcrafted features struggle to segment sky and non-sky regions effectively due to the complex and varied nature of these regions.

To address these challenges, we propose a dataset that contains outdoor images under a range of weather conditions, including all five weather types considered in this work. Importantly, we also provide ground truth sky masks for these images.

As depicted in Fig. 3.7, we employ a U-Net architecture to segment the sky from the image. A key difference between our approach and the traditional U-Net is the addition of a 3×3 tail convolution layer to the last convolutional layer, which refines the output. This tail convolution enhances the segmentation process, producing a foreground-background output where the sky is represented by white pixels and non-sky regions by black pixels.

The network performs five up and down-sampling operations, using a kernel size of 4×4, a stride of 2, and an input image size of 256×256. Multiple benchmark datasets have been used to train the SKYNET and have achieved high accuracy in segmenting sky regions across most weather conditions. However, in the case of haze, the color and texture of the sky closely resemble that of the haze, making it challenging to differentiate between the two, even for humans. As a result, the network's performance

is less accurate under hazy conditions.

Loss functions

We intend to determine the parameters of the network by minimizing a loss function, consisting of two different components.
Spatial Domain Loss: Standard L_2 loss function guides the network for the binary classification of pixels in the spatial domain and is defined as:

$$L_{sd} = \frac{1}{N}\sum_{i=1}^{N} \|x_i^{gt} - SN(x_i^{img})\|^2 \tag{3.2.6}$$

where $SN(I)$ is the SKYNET-based output for an input color image I, I^{gt} refer to the corresponding ground truth and N is the total number of images.
Threshold Loss: The input to SN is in the range of 0 to 1, but the output is a binary mask. The output from SN may result in some intermediate value or may even lead to overflow. To overcome the limitation, we use the sigmoid function in the tail of SN to limit the output to either 0 or 1:

$$O_i = S(SN(x_i^{img})) \tag{3.2.7}$$

where $S(.)$ represents the sigmoid function and O_i points to the mask output of the i_{th} sample image respectively. The proposed threshold loss may be formulated as:

$$L_{tl} = \frac{1}{N}\sum_{i=1}^{N} \log \|(O_i - SN(x_i^{img})\| + 1) \tag{3.2.8}$$

The overall loss function becomes:

$$L = \alpha L_{sd} + \beta L_{tl} \tag{3.2.9}$$

where α and β are weight parameters used to balance the two loss terms. We set both to 1 *i.e.* their default values to give equal weightage to both the components of the loss function.

Sky features

Sky color plays a significant role in predicting the weather conditions of an area. On a sunny day, the sky predominantly appears blue, as air molecules scatter blue light more than red light, leading to a relatively uniform blue color and lower av-

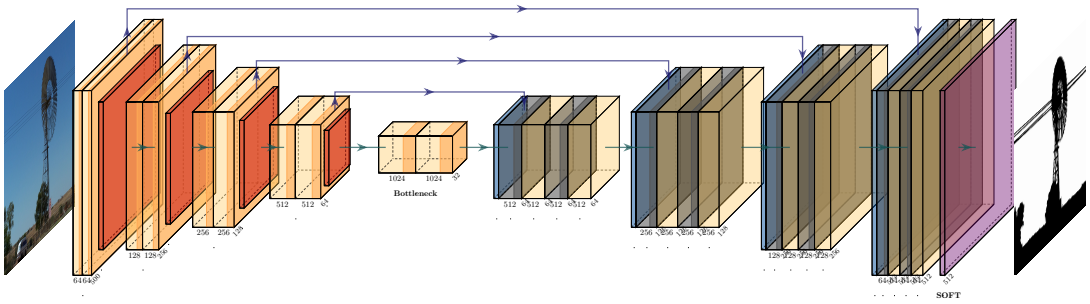


Figure 3.7: Sky Segmentation - SKYNET

Weather Type	*Texture*	*Color Intensity*	*Gradient*
Sunny	Smooth	Blue	Lower
Cloudy	Rough	Gray	Higher
Rainy	Rough	Gray	Higher
Winter	Transparent	White	Lower
Haze	Rough	Brown	Higher

Table 3.1: Sky Properties

erage gradient amplitude. In contrast, on cloudy or rainy days, the sky takes on a grayish-black hue due to the presence of clouds, which are primarily composed of water droplets. These droplets scatter a broader range of wavelengths of light, making the sky appear duller and resulting in a higher gradient amplitude. Furthermore, in sunny images, the absence of clouds or their scattered nature creates large smooth regions in the sky, further reducing the gradient.

During winter, the sky typically remains transparent and clear, with a color shift towards white. On the other hand, haze contributes a brownish granular tint to the sky, which leads to higher average gray intensity but a significantly lower gradient amplitude. This makes sky detection particularly challenging in hazy conditions, as the haze closely resembles the sky in appearance, complicating differentiation.

To capture the color information of the sky, we calculate the average of each of the red (R), green (G), and blue (B) color channels. These values are referred by f_{skyR}, f_{skyG}, and f_{skyB}, respectively. Additionally, to account for the texture of the sky, we compute the average gradient amplitude, which helps to assess the smoothness or roughness of the sky's appearance.

A summary of the sky properties observed during different weather conditions is provided in Table 3.1.

The gradient of an image gradient, for a channel c, is defined as,

$$f_{grad}(I_S) = \frac{\partial I_S}{\partial x}\hat{x} + \frac{\partial I_S}{\partial y}\hat{y} \tag{3.2.10}$$

where I_S stands for the sky region of the image, $\partial I_S/\partial$x, and $\partial I_S/\partial$y is the gradient in the X and Y direction, respectively. The mean gradient amplitude for a channel c is defined as,

$$f_{\nabla I_S,c} = \frac{1}{N_{sky}} \sum_{x,y \in Sky} \sqrt{(\frac{\partial I_S}{\partial x})^2 + (\frac{\partial I_S}{\partial y})^2} \tag{3.2.11}$$

Thus the sky feature f_{sky} can be expressed as:

$$f_{sky} = [f_{sky,c}, f_{\nabla I_S,c}] \quad c \in \{R, G, B\} \tag{3.2.12}$$

generating a 6-D sky feature vector.

Contrast Energy

Images captured in open environments exhibit varying levels of global and local contrast that correspond to different weather conditions. One important metric for quantifying local contrast in an image is the Contrast Energy (CE), which helps in capturing how much contrast exists within specific regions of the image. For this purpose, the segments derived from the non-sky areas of an image I are first identified using the Mean Shift algorithm.

To compute CE, the image is then decomposed using second-order Gaussian filters, which help in analyzing the contrast of each segment at different scales. Following this, all the filter responses are rectified, which involves adjusting the responses to ensure they are positive, and then normalized to ensure consistency across different segments. This step is followed by a non-linear contrast-gain control process, which further enhances the contrast. A thresholding step is applied to reduce noise in the image, ensuring that only significant contrast features are retained for further analysis.

The CE is computed separately for each of the components in the CIELab color space of the image. The CIELab color space is widely used in image processing because it separates the chromatic components (a* and b*) from the lightness component (L*), allowing for a more accurate measurement of contrast in terms of how humans perceive color. Specifically, the CE is computed for each of the components L, $a*$, and

$b*$ of the CIELab color space, as follows:

$$CE(I_c) = \frac{\alpha.Z(I_c)}{Z(I_c) + \alpha.\kappa} - \tau_c \tag{3.2.13}$$

$$Z(I_c) = \sqrt{(I_c \otimes h_h)^2 + (I_c \otimes h_v)^2} \tag{3.2.14}$$

where κ is a contrast gain set to 0.1 and τ_c is the noise threshold for each color channel c $\in$ [L, a*, b*], α is the maximum value of Z(I_c). The symbol $\otimes$ means convolution, while h_h and h_v are the horizontal and vertical second-order Gaussian functions using a 20$\times$20 filter window. The thresholds set for the L, *a, and *b channels are 0.2353, 0.2287 and 0.0528 respectively [70].

For each image, we select the normalized value of the largest twenty segments from the non-sky area as stated below:

$$f_{ce}(i) = w_i * CE_i(I) \tag{3.2.15}$$

$$w_i = N_i / N_I$$

Here, w_i represents the weight assigned to the contrast energy CE_i of the segment i of the image I. N_i denotes the number of pixels in segment i and N_I is the total number of pixels in the largest 20 segments of the non-sky area of image I. This process generates a 20-dimensional Contrast Energy feature vector f_{ce}.

Snow feature

Snow detection from an image is a challenging task due to the inherent characteristics of snow. The texture of snow is mainly light and soft, and in most, it appears uniformly white as it is an excellent light reflector. Moreover, snowflakes, driven by high winds, often scatter unpredictably, causing the trajectory of snowflakes to be disorderly. As a result, images taken during snowy conditions tend to look "noisy."

Since snow is predominantly uniform, continuous, and white, it leads to low saturation and very high gray intensity in the image. To detect its presence in an image, we focus on the lower 30% of the image, which is more likely to contain snow. However, the lower 30% of the image could also have areas with no snow, so the area is fragmented into multiple segments using the Mean Shift Algorithm [71].

For each generated segment S, we compute the mean normalized saturation. This helps in distinguishing snow from other regions based on its low saturation value and high gray intensity, as snow tends to exhibit a lack of vibrant color (low saturation)



Figure 3.8: Sky detection results for (a) cloudy image, (b) rainy image, (c) sunny image, (d) snowy image, and (e) hazy image.

but high brightness (high gray intensity). The normalized saturation for each segment is computed as follows:

$$f_{sat,S} = Mean\{\frac{S_i - \min(S)}{\max(S) - \min(S)}\} \quad pixel\ i \in S \tag{3.2.16}$$

Here $\max(S)$ and $\min(S)$ represent the maximum and minimum saturation of segment S respectively. This process generates a 20-dimensional saturation feature vector f_{sat} which is useful in distinguishing snow from other weather conditions.

3.2.3 CNN Model

A CNN model with the ResNet-101 backbone [72] is employed on a five-class image dataset to evaluate the performance of our proposed composite model, which integrates both hand-crafted and network-based extracted features. The loss function used is cross-entropy loss, and the CNN parameters are optimized by maximizing the average log probability of the correct label across the training examples.

3.2.4 Datasets and MLJSVM

Weather Datasets

Two Weather Dataset[60], providing Sunny and Cloudy weather images, Desnow Dataset[73] for snow images, and MWI Dataset[66] for Sunny, Hazy, Rainy, and Snow images have been used for training the model. In total, 30855 images were collated from all the mentioned benchmark datasets.

For particulate matter estimation, a large dataset with access to ground truth *i.e.* the actual values of $PM_{2.5}$ of the region, is essential for developing a reliable prediction model. The creation of this dataset is now described. The performance of the proposed models from different perspectives is evaluated in the ensuing subsections.

Sky Dataset and SKYNET

Segmentation of the object of interest from an image is a highly challenging task. Traditional segmentation techniques, such as threshold determination, often fail to provide satisfactory results. This limitation has led researchers to explore Deep Learning methods. However, even these methods struggle to produce optimal results when the data lacks a comprehensive representation of all the weather conditions under consideration.

As discussed in Section 3.2.2, a sky dataset has been proposed that includes all the weather conditions studied in this work. The dataset consists of 30,855 images, each of size 256×256, representing different weather conditions along with their corresponding ground truth. To create the dataset, we enlisted a helper to manually generate the mask for each image, where white pixels denote the sky area and black pixels indicate the non-sky area. Fig. 3.9 illustrates the weather images along with their corresponding sky masks.

For training the SKYNET model, we utilized 24,684 images, each of size 256×256, covering all weather conditions, with the remaining images reserved for testing. The proposed SKYNET was implemented on a PC with an Nvidia GeForce RTX 2060 Super. The batch size during training was set to 32, and the filters for each layer were initialized using He_normal. The ADAM optimizer was used to minimize the loss function defined in Eq. [3.2.9], with a learning rate of 0.0001. The total number of epochs was set to 200, with 100 iterations per epoch.

Multi-class Latent Joint Support Vector Machine (MLJSVM)

For each training image, we extract a 93-dimensional handcrafted weather feature and a 512-dimensional CNN feature. To train the SVM model, we randomly selected 80% of the data for training and reserved the remaining for testing. A cross-validation scheme was implemented for each round. We performed 10 rounds of training, recording the classification accuracy for each round, and computed the mean accuracy as the final result. A total of 24684 images were used for training, while 6171 images were used for testing the model.

We define the Multi-class Latent Joint Support Vector Machine (MLJSVM) as follows:

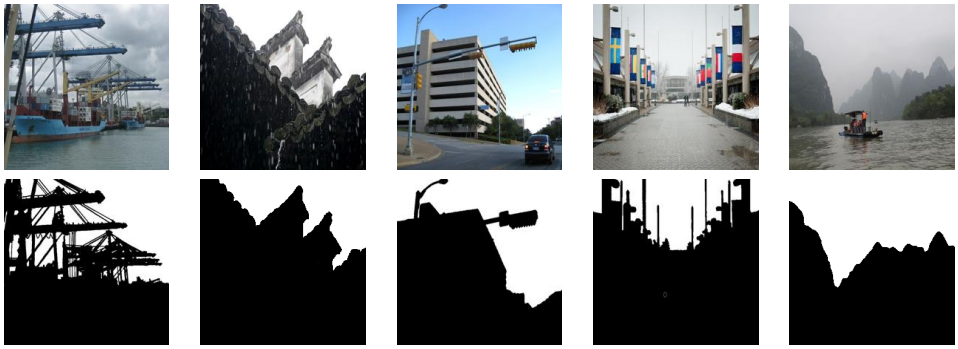


Figure 3.9: Sky Mask for (a) cloudy image (b) rainy image (c) sunny image (d) snow image (e) hazy image

Let $Y \equiv \{1, 2, \cdots, N\}$ be a set of N classes, where $N = 5$. We define ϕ, ψ and the block matrix W as

$$\begin{aligned} \psi(x_i, \beta) &= \beta x_i^T \in \mathbb{R}^{m \times d} \\ \phi(x_i, y, \beta) &= \begin{bmatrix} 0 & \cdots & 0 & \psi(x_i, \beta) & 0 & \cdots 0 \end{bmatrix}^T \in \mathbb{R}^{mN \times d} \\ W &= \begin{bmatrix} W_1^T & W_2^T & \cdots & W_N^T \end{bmatrix} \in \mathbb{R}^{mN \times d} \end{aligned}$$

where each W_y is a matrix containing m d-dimensional models and β is the m-dimensional latent variable. Here, $d = 605$, is the dimension of the input feature vector, and m is the number of training images. The prediction made by the algorithm is given by

$$\begin{aligned} y &= \psi(x_i, \beta) \\ y_i &\equiv \arg\max_{y \in Y} s_w(x_i, y), \\ s_w(x_i, y) &\equiv \max_{\beta \in \Omega_p} f_w(x_i, y, \beta), \\ f_w(x_i, y, \beta) &\equiv W.\phi(x_i, y, \beta) \end{aligned}$$

where Ω_p is defined as the positive p-norm unit ball, $\Omega_p \equiv \{\beta \in \mathbb{R}^m : \|\beta\|_p \leq 1, \beta_i \geq 0, \forall i = 1, \cdots, m\}$ and β can be interpreted as a model weight factor. The objective function is finally defined as:

$$\min_{W, \xi} \frac{\lambda}{2} \|W\|_F^2 + \sum_{i=1}^{n} \xi_i \tag{3.2.17}$$

such that

$$\begin{aligned} 1 - (s_w(x_i, y_i) - n_w(x_i, y_i)) &\leq \xi_i, i = 1, 2, \cdots, n \\ \xi_i &\geq 0, i = 1, \cdots, n \end{aligned}$$

where $\|.\|_F$ is the Frobenius norm, $n_w(x_i, y_i) \equiv \max_{y \in Y/\{y_i\}, \beta \in \Omega_p} f_W(x_i, y, \beta)$, n is the number of training images, λ is a regularization weight.

3.2.5 Analysis of experiments and results

The classification results have been reported under various parameter settings. The CNN classifier model was trained end-to-end on the datasets.

Individual Features

We utilized a Multi-class Latent Joint Support Vector Machine (MLJSVM) to evaluate the individual features for weather classification. The classification results are summarized in Table 3.2. The Convolutional Neural Network (CNN) performance is sensitive to the choice of hyperparameters, which were set as follows: momentum = 0.9, decay = 0.9, and learning rate = 0.01. As mentioned, we trained the CNN model using a dataset of 24,684 images from various weather conditions, with 6,171 reserved for testing. For fusion of weather and CNN based features, we performed simple concatenation operation to generate the feature vector corresponding to each image.

The results show the accuracy for each feature and the performance when each feature is omitted one at a time. Unsurprisingly, visibility and sky-based features emerged as the most important for weather classification. This can be attributed to the fact that the sky behaves distinctively under different weather conditions, and its color variation is more consistent with the non-sky regions. Additionally, visibility impacts how an image appears under varying weather conditions, highlighting its importance in classification tasks.

However, there were some failure cases, particularly with snow and rainy images. In snow images, the sky often resembles a transparent, clear, white sunny day, which can mislead the classifier. Additionally, the snow in certain images may be too minimal to be detected properly. On the other hand, detecting rain streaks in a static image remains a highly challenging task, which led to errors in classifying rainy weather images.

Ablation Study

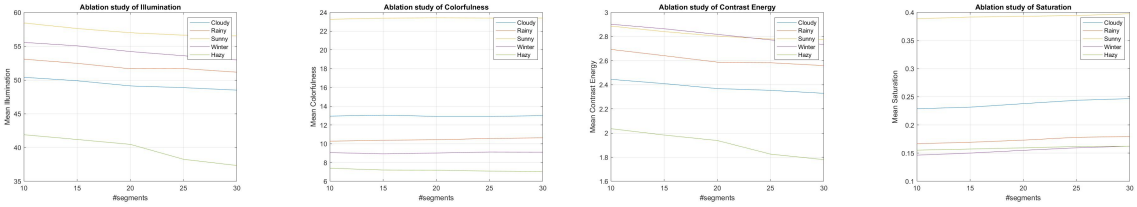


Figure 3.10: Ablation study of segment numbers with respect to the extracted feature.

We conducted an ablation study to assess the performance of our system when each of the features is excluded. The results in Table 3.3, demonstrate that the extracted handcrafted weather features contribute significantly to the overall weather classification. Excluding any one of them leads to a noticeable drop in accuracy. Notably, while the CNN-based feature is important, the combination of all handcrafted features outperforms the CNN-based model classification accuracy.

In addition, we performed another ablation study to analyze the impact of the number of segments used for each handcrafted feature. This study was conducted on a subset of 500 images from each of the five weather conditions. We evaluated how the average feature values change with the number of segments ranging from 10 to 30. The results, shown in Fig. 3.10, indicate that for features like Colorfulness and Saturation, there is minimal variation in the average feature values regardless of the number of segments used.

However, for features such as Illumination and Contrast Energy, we observed that the average values of these features decrease as the number of segments increases. Beyond 20 segments, there is a significant drop in the mean value, which suggests that additional segments with lower illumination and contrast energy are being incorporated into the feature vector. Based on the observations, we use 20 segments for each feature to maintain optimal performance.

Comparison of performance

We present the performance of our overall classification model and compare it with other state-of-the-art weather detection systems and CNN classifiers. Our findings show that relying solely on the CNN model without incorporating weather-specific features results in a performance drop. While neural networks are powerful tools for encoding scene structure and image characteristics, they cannot effectively capture the unique weather-related information needed for accurate weather classification.

To address this, we implement the Multi-class Latent Joint Support Vector Machine (MLJSVM) directly on the 605-D feature vector, which consists of 93 handcrafted

Feature	*Accuracy*
Illumination	34.3
Colorfulness	41.8
Sky	73.0
Visibility	83.0
Contrast Energy	35.2
Saturation	39.7
CNN Model	73.15

Table 3.2: Classification results obtained using only the specified features are presented

Feature	*Accuracy*
Illumination	89.0
Colorfulness	90.0
Sky	73.0
Visibility	74.0
Contrast Energy	90.0
Saturation	88.7

Table 3.3: Classification results obtained by omitting the specified feature and the features extracted from ResNet-101 are presented with all figures in %.

Model	*Accuracy*
Ours	94.19%
Li[74]	83.79%
Resnet-101	73.15%

Table 3.4: Classification results of different methods.

features and 512-D features extracted from the CNN model. We experimented with different kernels, including Gaussian, RBF, and polynomial kernels in linear and non-linear forms. The kernel that delivered the best performance was the linear kernel, with the parameter value $C = 1$ providing the optimal result within the range $C \in [0.0001, 100]$.

The primary misclassification occurred with Snow and Rainy images, as these weather conditions lack distinctive features that can effectively differentiate them from other conditions. The CNN model's hyperparameters were fine-tuned, with the best results achieved when the learning rate, momentum, and weight decay were set to 0.0001, 0.9, and 0.005, respectively. These optimized parameters allowed the model to perform better, though challenges remained in accurately classifying snow and rain images due to the absence of distinctive weather-specific features. Not much research has been conducted on the five-weather classification problem considered here. Most existing studies have focused on classifying just two or three weather types, primarily Sunny, Cloudy, and Haze. In [74], images with five different weather conditions—Sunny, Cloudy, Haze, Snow, and Rainy—are classified by fusing various weather features with different neural networks, with the highest accuracy achieved using ResNet-50. A performance comparison is presented in Table 3.4.

It is important to note that Liu in [60] focused solely on Sunny and Cloudy weather classification. We have tested our approach on the same classes using publicly available code, as implementing state-of-the-art works can be challenging. The accuracy reported in [60] was 91.4%, while our approach achieved an accuracy of 97.63%.

To further enhance the effectiveness of our work, we incorporated an end-to-end CNN classifier for multi-weather classification. We used the state-of-the-art CNN model, ResNet-101, which was fine-tuned to achieve the best results for our specific needs.

To ensure fairness, we maintained a consistent set of parameters and performed fine-tuning by varying hyperparameters such as learning rate, momentum, and weight decay. The combined use of CNN and non-CNN features led to a significant accuracy improvement, from 73.15% to 94.19%, demonstrating the effectiveness of the weather-specific handcrafted features.

3.3 Image-based Determination of Particulate Matter Concentration

Advancements in the world economy and changes in people's lifestyles have led to extensive use of fossil fuels in power generation, transportation, and energy-intensive industries, leading to severe air pollution worldwide. Air pollution is one of the gravest challenges humans face, causing serious risks to public health, especially for those living in China and India, besides playing a crucial role in global warming and consequent climate change. $PM_{2.5}$ and PM_{10} (particulate matter with a diameter less than 2.5μm and 10μm, respectively) are used to measure the concentration of tiny and fine particles in the air [22].

Different approaches exist for determining PM concentration. The Central Pollution Control Board (CPCB), a body constituted by the Government of India, has prescribed three different methods for measuring the density of $PM_{2.5}$ and PM_{10} vide its notification dated 18th November 2009. These are the Gravimetric method, Beta Attenuation method (BAM), and Tapered Element Oscillating Microbalance (TEOM), of which automatic monitoring stations follow BAM. It works by measuring the absorption of beta radiation by solid particles collected from the air. The reference PM data for this article has been obtained from such a station.

The gravimetric method determines particulate matter (PM) concentration by measuring the weight difference of filters before and after sampling. Accurate results rely on minimizing electrostatic effects, controlling moisture absorption, maintaining stable temperatures, and careful sample handling. This method is endorsed by ASTM International standards D2274 and D2276 for measuring PM emissions in petroleum products and aviation fuels.

Beta attenuation monitoring (BAM) is another widely used technique for estimating PM_{10} and $PM_{2.5}$ concentrations. It works by measuring the absorption of beta radiation by solid particles collected from the air.

Optical instruments for particle measurement operate on the principle of light scattering. When particles pass through a beam of light, some light is scattered, and detecting this scattered light allows for particle counting. Additionally, the intensity of the scattered light is related to the size of the particles, enabling measurements of particle size beyond just counting.

Since constructing monitoring stations at short distances is financially not viable, their locations are determined such that land, power, and financial constraints are optimized. With an area of approximately 200 Km^2, 22 monitoring stations, includ-

ing seven automatic stations, monitor the air quality of the city. Moving beyond the air sensors, an image-based approach for estimating the density of particulate matter forms a low-cost alternative for addressing such a challenge, with the availability of affordable cell phones with high-resolution cameras.

3.3.1 Related Work

In recent research, attempts have been made to estimate $PM_{2.5}$ levels from images. Some works, such as Gu et al. [75], have built naturalness statistics models based on entropy features in both spatial and temporal domains, followed by non-linear mapping functions. Others, like Lieu et al. [76], have explored contrast, transmittance, and entropy metrics, focusing on a region of interest (RoI) to improve prediction performance using support vector regression (SVR). Similarly, Hsieh et al. [77] introduced RoI selection based on high-frequency information and the use of Improved Dark Channel Prior (IDCP) for haze modeling, combined with Relative Humidity (RH) data to train an SVR model.

Chakma et al.[78] introduced a deep convolutional neural network (CNN) model to classify natural images into different categories based on their $PM_{2.5}$ concentrations. They proposed a dataset consisting of 591 daytime images from Beijing with corresponding $PM_{2.5}$ concentrations for training the model. Nitin et al.[79] proposed an IoT-based real-time air quality index (AQI) estimation technique using images and weather sensors on Indian roads. It used a fusion of traffic density, visibility extracted from images, temperature, and humidity from sensors, with a Machine Learning (ML) model trained on those features to predict the AQI. The authors also proposed a dataset containing 5048 images and co-located AQI values across different seasons by driving on the roads of Hyderabad city. However, these methods left room for improvement in terms of performance metrics like R^2, and many also depended on access to atmospheric data or location-specific information, which posed challenges.

3.3.2 Formulation of the approach

The method for particulate estimation consists of the following steps:

1. Manual selection of RoIs based upon relative distance from the camera (one near object and one far object). In the case of entire images, it consists of both near and far objects.

2. Dehazing the image using Pyramid Fusion and computing SSIM between the

original input and the dehazed image (either for the respective RoIs in the case of the first approach or for the entire image as in the second one).

3. Training Regression models using the SSIM and day, month, and hour as input features.

Fig. 3.11 presents these in a block diagram for easy understanding.

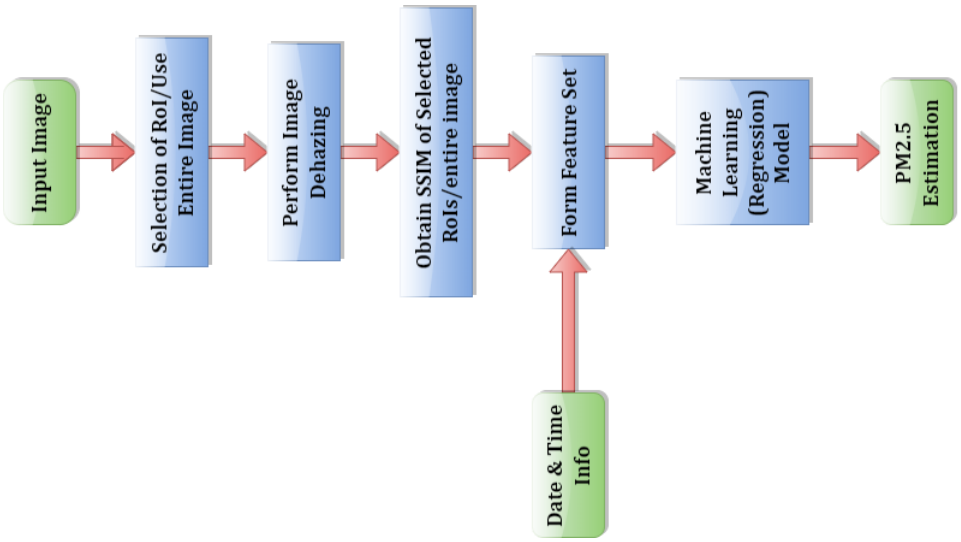


Figure 3.11: Block diagram of the proposed method.

Each of the steps is now elaborated upon.

Selection of RoIs

It has been observed in [76] that focusing on a RoI improves the performance of the ensuing prediction model. While observing the visibility of objects at a distance from a fixed point, in this case, the camera's location, the variation in visibility can be noticed between a near object and a relatively far object. As seen in Fig. 3.12, high $PM_{2.5}$ concentration changes the visibility of the objects, due to haze formation. It can also be observed that the changes in visibility are more for distant objects as compared to nearer objects, to the extent that the distant objects become invisible during high $PM_{2.5}$ concentration. Therefore, the two RoIs are manually selected because the difference in their visibility can be used as an indicator of the $PM_{2.5}$ concentration. Fig. 3.13 illustrates a judicious choice of the two RoIs as required by the first approach. For the second approach, utilizing the entire image, the images were selected by ensuring that they contain both near and distant objects within their view, for pixel-wise comparison. While selecting the RoI's from an image, importance is given to objects that

are fixed, having substantial height and width, not very far such that it's visibility is not lost, and are present mainly in central area of the image. Further, with regard to camera calibration, the focus and resolution has been set fixed for different scenarios such that the objects that are intended to be covered in the images are clearly visible during high visibility during low particulate concentration. This has been done so as to do comparative analysis of object's visibility during high particulate concentration.

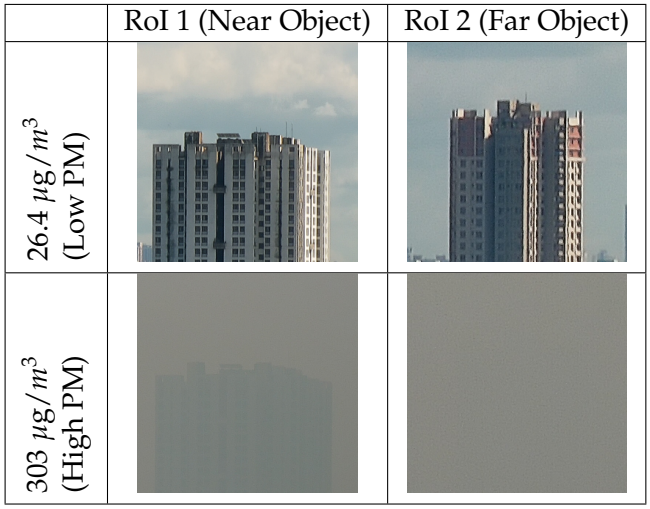


Figure 3.12: Visibility of selected RoIs with respect to $PM_{2.5}$ concentration.

Image Dehazing for creating a Reference Image

Since a reference, clear image (i.e. free from the effects of $PM_{2.5}$) is not available in practice, the proposed techniques attempt to construct such an image for reference so that the visibility and its loss may be quantified. For this purpose, the Structural Similarity Index Measure (SSIM) is employed as a full-reference metric after dehazing the input image. For dehazing, the Pyramid Fusion algorithm in [25] has been adopted which does not require any prior information about the hazy image or any atmospheric parameter.

After dehazing the image, following the first approach, the mean SSIM between the original image and the dehazed image is computed on the two RoIs, as shown in Fig. 3.14. This metric is used to quantify the visibility, which varies depending on the thickness of the haze, in turn depending on the $PM_{2.5}$ concentration level and also according to the distance of the object from the camera, as demonstrated in Fig. 3.14. For the second approach, the SSIM of the entire image is computed as an SSIM feature



Figure 3.13: Selection of RoIs based on the relative distance of objects (one near and one relatively far)

map of the same dimension as the input image, henceforth referred to as SSMAP. Fig. 3.15 shows two images of the same view, corresponding to two different levels of $PM_{2.5}$, the dehazed images and SSMAP. Fig. 3.16 shows images with different views in the multi-view dataset and their corresponding SSMAPs.

Training Regression Models

It is well known that weather/meteorological conditions, such as very high relative humidity, lead to haze formation. Fig. 3.17, which is a plot of the quarter-hourly $PM_{2.5}$ measurements obtained from the air-quality monitoring station at Rabindra Bharati University (described later in Section 3.3.3), shows that the $PM_{2.5}$ concentration also varies depending on not only the season but also the time of the day. As shown in the figure, there is a massive increase in $PM_{2.5}$ concentration during the months from November to March, as compared to the rest of the year. Therefore, a convenient way of including this feature is to consider day, month, and hour as features for training in addition to SSIM. The day, month, and hour variables are cyclically encoded as shown in Eq. [3.3.1-3.3.2] to represent them as periodic variables,

$$x_{sin} = \sin(\frac{2\pi x}{x_{max}}) \tag{3.3.1}$$

$$x_{cos} = \cos(\frac{2\pi x}{x_{max}}) \tag{3.3.2}$$

The input features, i.e., SSIMs, day, month, and hour, are used for training three regression models, against $PM_{2.5}$ concentration measured from the monitoring station, treated as ground truth. The models are built by training the Support Vector Regressor

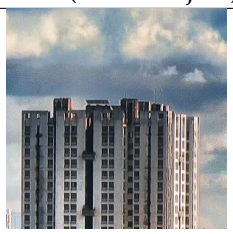
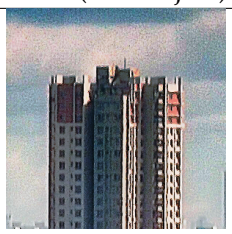
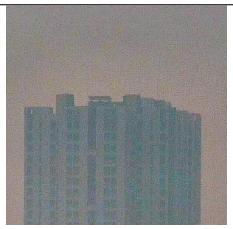
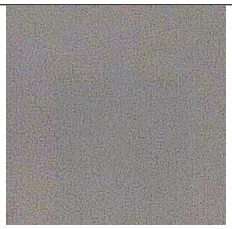
	RoI 1 (Near Object)	RoI 2 (Far Object)
26.4 $\mu g/m^3$ (Low PM)		
SSIM	0.8270	0.6640
303 $\mu g/m^3$ (High PM)		
SSIM	0.6372	0.4732

Figure 3.14: Visibility and SSIM of selected RoIs after dehazing.

(SVR), Random Forest Regressor (RFR), and Extra Trees Regressor (ETR) with test R^2 = 0.79, 0.92, and 0.93, respectively, for building the RoI-based models of the first approach. Similarly, for the second approach, the resultant SSIM feature map along with day, month, and hour are trained with ETR on a single view and multi-view datasets, resulting in R^2 = 0.87 and 0.8, respectively on test datasets.

3.3.3 Experimental results for Particulate Matter Concentration system

Particulate Matter Datasets

The acquisition of the dataset covering outdoor images with multiple views has been prepared for task accomplishment. Since low light affects the visibility of nighttime images, an accurate estimation of $PM_{2.5}$ and PM_{10} concentration requires high-resolution images. We have used the Dahua PTZ network camera DH-SD59430U-HNI, having a lens with 4MP resolution, size 2592×1520, placed on the terrace above the ninth floor of S.N. Bose Bhavan of the Indian Statistical Institute, Kolkata, at an approximate height of 120 feet. The reference PM data were obtained from the Air Quality Monitoring Station at Rabindra Bharati University, Kolkata, at a distance of approximately 2.5 km from various locations captured by the camera. Six views are added in the camera settings to cover structures at varying distances or horizontal depths. Since the monitoring station measures the concentration levels at an interval

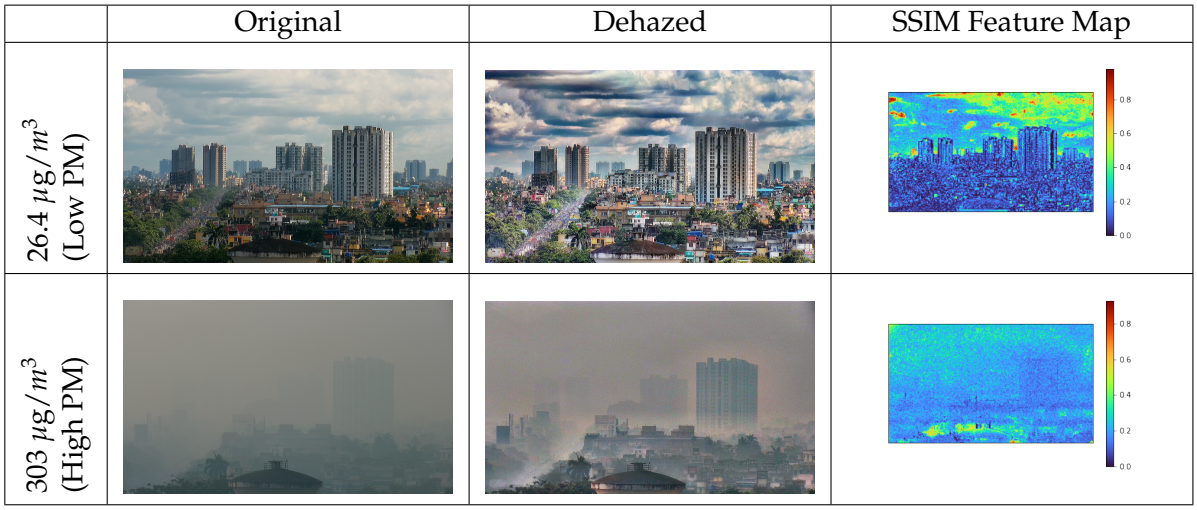


Figure 3.15: Images along with dehazed image and SSIM feature map of low and high $PM_{2.5}$ respectively.

of 15 minutes, the same strategy has been used for image acquisition. Thus, each view is taken at an interval of 15 minutes, i.e., the time difference between two consecutive images with different views is 150 seconds. The location of the camera and the corresponding Air Quality Monitoring station are shown in Fig. 3.18. The figure also shows the spatial distribution of other monitoring stations in Kolkata.

From images collected during the period from June 2021 to January 2023, a total of 3620 images were curated for training and 905 images for testing. The images with one particular view were used in the first approach for convenient selection of RoIs, as shown in Fig. 3.13. Another multi-view dataset was also curated, consisting of 3375 images for training, and 844 images for testing, from which two images were shown in Fig. 3.16. For ground truth, quarter-hourly measurement data of $PM_{2.5}$ recorded at the Air Quality Monitoring Station, located at Rabindra Bharati University (RBU), Kolkata, was collected, as the image capture location falls within a 2.5 km radius of the monitoring station. The open-source data is available on the official website of the Central Pollution Control Board of India.

Evaluation of RoI Based Model

Evaluation of the trained models is performed on the test dataset, with Coefficient of Determination (R^2) along with Pearson's Coefficient (PCC), Kendall's Rank Coefficient (KRC), and Root Mean Square Error (RMSE) being used as the evaluation metrics. The results of the RoI-based models are given in Table 3.5, and the scatter plot

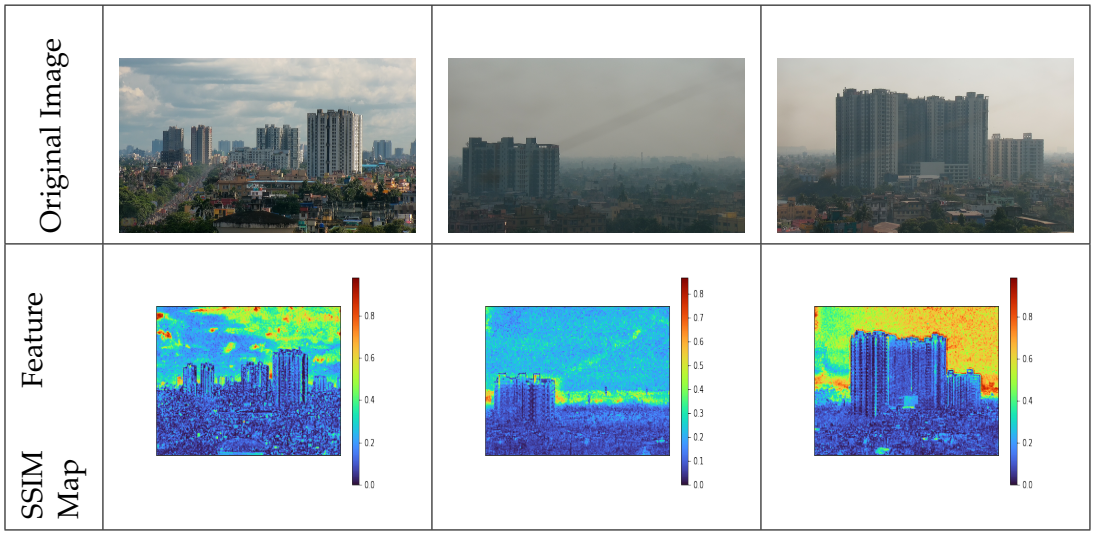


Figure 3.16: Images and their corresponding SSIM feature map from the multi-view dataset.

of predicted $PM_{2.5}$ concentration against measured $PM_{2.5}$ concentration is provided in Fig. 3.20(a),(b),(c). The result shows that ETR performs best out of the three RoI base models with $R^2 = 0.93$.

Evaluation of Entire Image Based Model

Since ETR gave the best result for the RoI-based model, it was chosen as the Regression model for the entire image-based estimator, trained on the single-view dataset and multi-view dataset separately. The results are given in Table 3.6, and the scatter plot of predicted $PM_{2.5}$ concentration against measured $PM_{2.5}$ concentration is provided in Fig. 3.20(d) and (e) respectively. From the results, it is evident that the RoI-based model performs significantly better than the entire image-based model. Of the two datasets, the model performs slightly better on the single-view dataset with $R^2 = 0.87$.

Model	R^2		**PCC**		**KRC**		**RMSE**	
	PF	***CAP***	***PF***	***CAP***	***PF***	***CAP***	***PF***	***CAP***
SVR	0.79	0.89	0.69	22.01	0.75	0.87	0.63	23.95
RFR	0.92	0.96	0.87	13.83	0.88	0.94	0.85	16.23
ETR	**0.93**	**0.97**	**0.9**	**12.56**	0.9	0.95	0.88	15.2

Table 3.5: Results of RoI-Based Models as per Dehazing Methods: Pyramid Fusion(PF) and Colour Attenuation Prior(CAP).

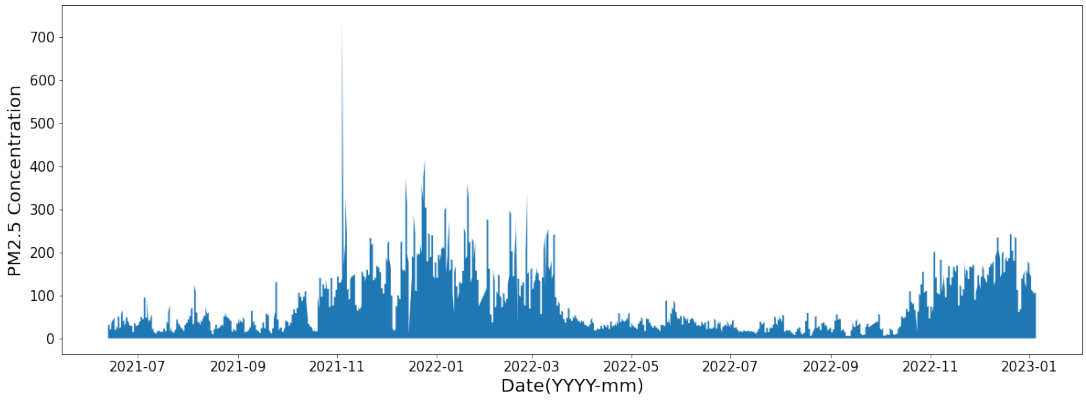


Figure 3.17: Variation of $PM_{2.5}$ concentration from Jul-2021 to Jan-2023.

Dataset	R^2		**PCC**		**KRC**		**RMSE**	
	PF	***CAP***	***PF***	***CAP***	***PF***	***CAP***	***PF***	***CAP***
Single-View Dataset	0.87	0.94	0.81	16.9	0.85	0.93	0.79	18.17
Multi-View Dataset	0.8	0.89	0.75	19.97	0.77	0.88	0.73	21.39

Table 3.6: Results of Entire Image-Based Models based on ETR for different Dehazing Methods: Pyramid Fusion(PF) and Colour Attenuation Prior(CAP).

Comparison of Prediction based on Dehazing Method

The performance of the proposed approaches was compared by applying a different dehazing algorithm. For this purpose, a single image dehazing method using Colour Attenuation Prior (CAP), as proposed by Zhu et al [80], was selected. Table 3.5 shows the comparison of the metrics, R^2, PCC, KRC, and RMSE of the RoI-based models between the dehazing method in our proposed approach, i.e., Pyramid Fusion, and CAP, where the former shows better performance than the latter. A similar comparison with the entire image-based method was performed. As shown in Table 3.6, the method applying Pyramid Fusion outperforms that of CAP, regarding all the metrics computed.

Comparison of PM2.5 Prediction Method

The proposed prediction method was compared with the Picture-based Prediction of $PM_{2.5}$ Concentration (PPPC) model as per Gu et al [75]. Since PPPC is not a regression model, the performance cannot be compared using R^2. Therefore, Pearson's Cor-

relation Coefficient (PCC), Kendall's Rank-order Correlation (KRC), and Root Mean Square Error (RMSE) have been calculated, which are the metrics used in [75]. The PPPC model was implemented to predict $PM_{2.5}$ on the test datasets. Table 3.7 provides the comparison between our proposed methods and the PPPC method on Single-view Datasets and Multi-view Datasets. The proposed methods are observed to perform far better than the PPPC model.

Comparison of Results on a Different Dataset

To validate our proposed method on a different geographic location, a dataset of images from Shanghai, China was selected, which was used for $PM_{2.5}$ prediction by Liu et al[76]. The dataset contains 1882 images, of which 1505 were used for training and the rest for testing. Fig. 3.19 demonstrates a viable choice for RoI selection. Since RoI-based ETR yielded the best result for our original dataset, it was also implemented here. Table 3.7 shows that, comparing the proposed method's results on the Shanghai dataset with the best outcomes in [76], the former performs better.

Dataset	Method	PCC	KRC	RMSE
Single Dataset	PPPC	0.48	0.5	139.13
	SVR	0.89	0.69	22.01
	RFR	0.96	0.87	13.83
	ETR (ROI based)	**0.97**	**0.9**	**12.56**
	ETR(Entire Image)	0.94	0.81	16.9
Multi Dataset	PPPC	0.47	0.43	134.86
	ETR(Entire Image)	**0.89**	**0.75**	**19.97**

Dataset	Method	R^2	RMSE
Shanghai	Liu et al (PCA-SVR)	0.76	13.76
	Liu et al (SBFS-SVR)	0.76	13.65
	ETR (ROI based)	**0.87**	**10.24**

Table 3.7: Comparison of performance of the proposed methods with PPPC model on Single-View, Multi-View, and Shanghai Datasets.

3.3.4 Concluding remarks

Image-extracted features, inspired by the human visual system, have been employed for weather detection. Features based on illumination detection are employed to identify sunlight in an image. Additionally, features that evaluate the visibility of a scene have been introduced.

The sky may not always be present in every image, meaning some images might lack sky-related features. This could lead to errors in the classifier, as it typically requires all features to be present simultaneously. A homogeneous clustering approach has been used to address this issue, where images are grouped into distinct clusters. Features extracted from images within the same cluster are homogeneous, ensuring consistency and improving classification performance. Furthermore, the proposed sky detection model, SKYNET, segments the sky from the rest of the image, while a robust sky-mask dataset has been designed to train the SKYNET model.

The final classification is achieved by combining handcrafted features with 512-dimensional feature vectors extracted using ResNet-101 and a Multi-class Latent Joint Support Vector Machine (MLJSVM) classifier to accurately categorize images into five weather conditions: “Sunny”, “Cloudy”, “Haze”, “Snow”, and “Rainy”.

Since the concentration of particulate matter is heavily influenced by the prevailing weather conditions, the output of the system for multiweather detection is likely to be a valuable input for particulate concentration determination. Further performance improvement would be explored along these lines. As the RoI based approach is observed to perform slightly better than the approach using the entire image, automatic selection of the RoI is also worth looking into. An important limitation of the methodology adopted remains: it fails for night-time images. In the absence of a strong, ambient illumination, the acquired images are granular, having artifacts which are liable to be mistaken for particulate matter, and also suffer from the effect of ’spurious’ light sources. An approach, employing deep learning, for dealing with night-time images, forms the subject of the next chapter.

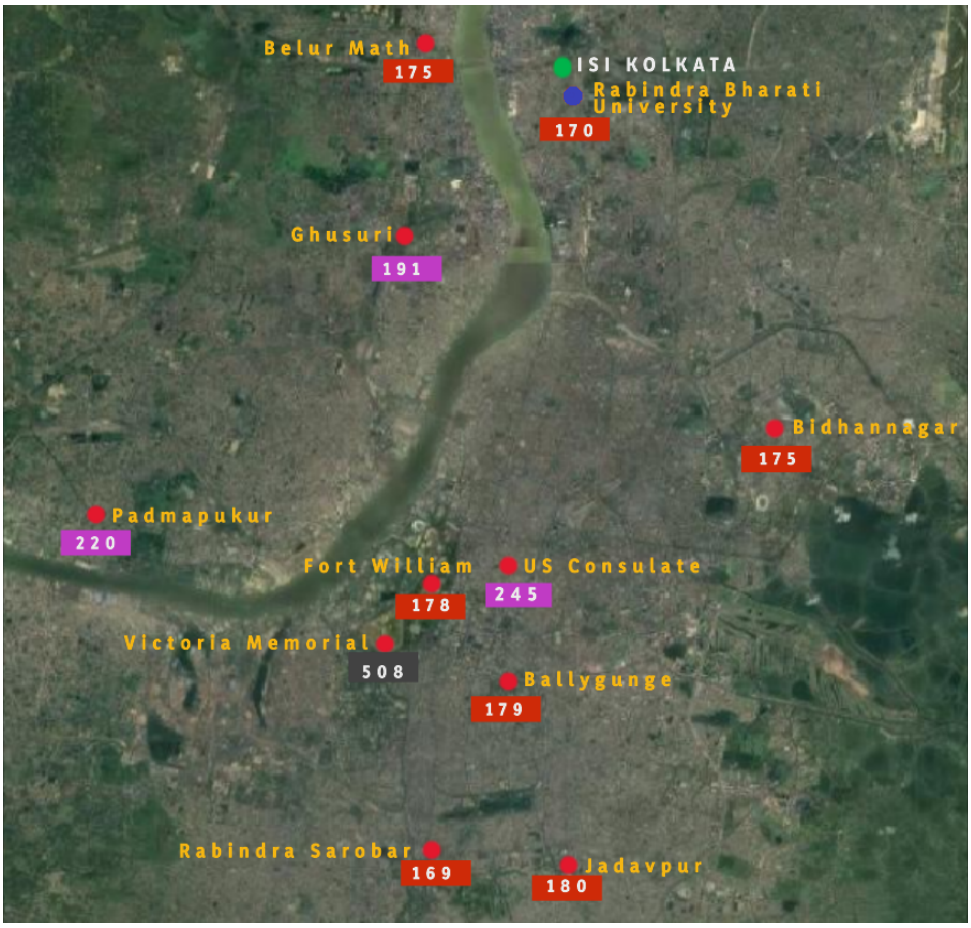


Figure 3.18: The distribution of air quality monitoring stations in and around Kolkata with a green dot indicating the location of ISI, Kolkata. The values shown in the image pertain to the AQI level recorded at the monitoring station on 5th February 2025 at 15:00. The station shown with the blue dot is the location of Rabindra Bharati University used for ground-truth measurement. Source: `https://aqicn.org/map/kolkata/`

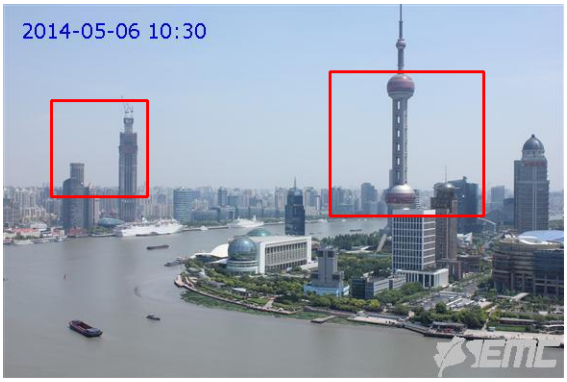
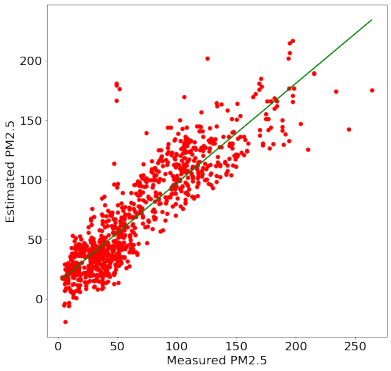
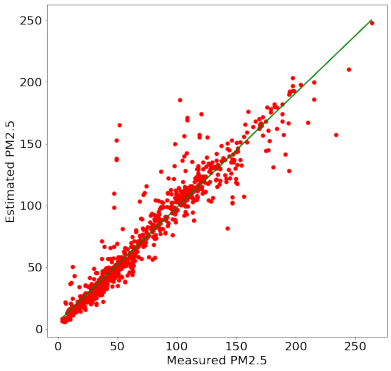


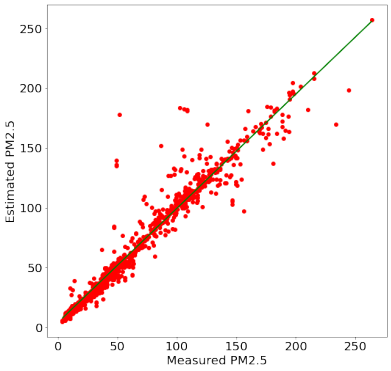
Figure 3.19: Selection of RoIs based on the relative distance of objects (one near and one relatively far) on Shanghai images.



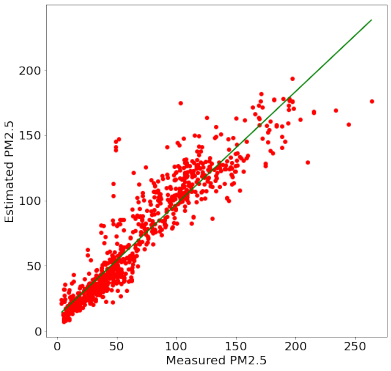
(a)



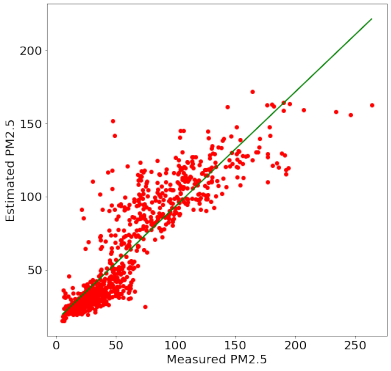
(b)



(c)



(d)



(e)

Figure 3.20: Scatter plots (Estimated $PM_{2.5}$ vs Predicted $PM_{2.5}$) of RoI-based Models a) SVR, b) RFR, and c) ETR, and entire image-based Models d) ETR on Single-View dataset, e) ETR model on Multi-View Dataset.

Chapter 4

Estimation of Particulate Matter Concentration from night scenes: From REAL-Net to Dark-AirNet

4.1 Introduction

Air Quality Index (AQI), representing the severity of pollution for ordinary people, is computed by various Governments by considering a weighted average of various pollutants present in the air. Government regulators in many countries monitor the air quality of industrial areas to ensure they comply with the prevalent environmental standards. Many high-polluting industries, such as iron & steel, petrochemical refineries, etc., use sprinklers at regular intervals of the day to reduce the concentration of particulate matter in the air. Lack of a low-cost measuring system for air pollution forces them to deploy highly costly and bulky machines to realize the objective. Particulate matters, especially $PM_{2.5}$ and PM_{10}, are predominantly given the highest weightage in AQI as these can penetrate deep into our lungs, causing diseases like bronchitis, asthma, and even cardiovascular ailments. According to the Global Burden of Disease from Major Air Pollution Sources: A Global Approach by the Health Effects Institute (HEI), India recorded 8,67,000 deaths in 2017 due to exposure to fine particulate matter.

Han et al.[81] show that air quality in many metropolitan areas at nighttime is poorer compared to daytime. Nighttime images have very low intensity as compared to daytime due to poor lighting conditions. They are captured under artificial lighting conditions, where there is a lack of natural light. Further, severe weather conditions

and steep variations in meteorological factors such as temperature, humidity, wind speed, etc., may make sharp changes in the concentration of the particulate matter. These factors often adversely affect image quality and can obscure distant objects. Daytime images are typically captured in natural sunlight, which provides strong lighting and high visibility of objects. The absence of this at night makes the extraction of key features from nighttime images challenging.

The advancement of artificial intelligence research enables the precise estimation of air pollution with an image-based approach, despite the hurdles. The absence of a strong ambient illumination in outdoor images acquired at night, along with ensuing problems as mentioned earlier, necessitates that such images be treated differently from daytime images. Developing a single model to tackle both day and night scenarios is not only an arduous task, but the performance is also expected to be far from optimal since the features of nighttime images differ dramatically from those acquired during the day. Hence, the analysis of nighttime images for air pollution estimation needs special attention, employing deep learning networks.

Many estimation studies use machine learning-based models such as Support Vector Machine (SVM)[76], multi-kernel learning[82]. However, these feature-based models are highly sensitive to the features extracted. Their performance can vary drastically based on changes in environmental conditions. Since image-based features are extracted manually to estimate the concentration of particulate matter, machine-learning methods are less robust than deep learning-based methods. Further advancements in deep learning techniques are expected to increase the accuracy of particulate estimation.

We introduce an encoder architecture that is deemed appropriate since our target output is the PM value and not an image, which would be the output of the usual encoder-decoder architecture. Night images contain regions such as edges and boundaries of various structural objects, as well as the spectral effect introduced by the particulate matter due to light illumination, such as flash or flood light. This motivates us to look for a solution to extract such high and low-frequency aware features. Further, buildings at far distances exhibit low visibility, whereas close structures have higher visual clarity. Low particulate matter concentration results in clear visibility, whereas high concentration causes low visibility. To avoid being confounded by these effects, the use of depth information is essential in making accurate estimations of the concentration.

Proper training of the model requires robust data, which includes images acting as input to the model and the actual concentration values from the air quality monitor-

ing station closest to the site of image acquisition. The model learns parameters while training, minimizing the loss computed between the actual and predicted concentration values. The training process is continuous, updating the learnable parameter values until the loss reaches an acceptable level. The particles responsible for the $PM_{2.5}$ and PM_{10} can have multiple sources, both natural and anthropogenic. This work considers a primarily urban area that is dominated by fumes from vehicular emissions, dust emanating from roads and construction activities, and road dust. Monitoring stations are located mostly in urban areas, thus placing a constraint on the availability of proper reference data and hence restricting our work.

Section 4.2 provides a literature survey, reviewing related works in the field of image-based pollution estimation and air quality monitoring. Section 4.3 provides detailed information about the dataset developed by us to create and train the model. Section 4.4 elaborates on the architecture and methodology of the proposed model, including key innovations and technical details. Section 4.5 finally presents the experimental setup and evaluation metrics used to assess the performance of the model. It also includes an ablation study to justify the design choices made in the model.

4.2 Related Work

This section provides a comprehensive overview of the existing literature by categorizing existing research in the area into two major approaches for estimating the concentration of $PM_{2.5}$ and PM_{10} and that of depth estimation.

4.2.1 Machine Learning based PM concentration estimation

The reviewed studies highlight various methods used for estimating $PM_{2.5}$ and PM_{10} concentration from images, with most focusing on leveraging image-based features and machine-learning techniques. Visibility and light scattering, both significantly influenced by particulate concentration, directly affect the quality of images, causing haze [83] and reduced clarity [84], which complicates the estimation process. Several methods have attempted to address this challenge. Liu et al. [76] used image features like contrast and entropy along with Support Vector Regression (SVR) to estimate particulate levels, but this approach struggles to capture the complex relationships between image features and particulate matter, especially under varying environmental conditions. Zhang et al. [82] employed multi-kernel learning with image-based features to predict the Air Quality Index (AQI), but the model's performance is limited

by its inability to adapt to changing environmental conditions. Yang et al. [85] combined the Dark Channel Prior (DCP)[86] method with transmission models based on relative humidity to estimate particulate concentrations, though their approach still depends on factors like humidity that can vary greatly. Liu et al. [87] used a combination of spatial contrast, DCP, and HSI color differences, but like others, their method faces challenges in adapting to changing environmental factors. Yue et al. [88] integrated structural and color information for particulate estimation, but their method may overlook important dynamic aspects of air quality. Hsieh et al. [77] improved DCP with high-frequency information from Regions of Interest (RoI), yet their model still faces limitations due to its reliance on external factors like lighting and humidity. Finally, Chen et al. [89] proposed a high-resolution dataset with detailed pollution data and images, but even with this data, their model is only partially able to address the complexities of particulate estimation under varied conditions.

These methods all share a common set of limitations. The features used for $PM_{2.5}$ estimation are often insufficient to fully capture the complexity of particulate matter occurrences and fluctuations, leading to high errors in predictions. Additionally, many models are not adaptable to changes in meteorological conditions, such as temperature and pressure, which are crucial for accurate estimation. Finally, the issue of light intensity, particularly in low-light conditions like during nighttime, further complicates the task. These challenges highlight the need for more robust, adaptive techniques, particularly those that can handle a broader range of environmental factors and leverage deep learning to learn complex features directly from raw data.

4.2.2 Deep Learning based concentration estimation

Advancements in high-speed GPUs and the development of deep learning architectures have significantly improved the performance of air pollution estimation models, particularly in comparison to traditional machine learning methods. Convolutional Neural Networks (CNNs), with their ability to efficiently extract spatial features, enable direct analysis of images without the need for manual feature extraction. Several studies have employed CNNs for estimating particulate matter concentration. For instance, Li et al. [90] combined dark channel and haze-image models with Deep Convolutional Neural Fields [91] to estimate the scattering coefficient and depth, followed by non-linear mapping to predict $PM_{2.5}$ levels. Zhang et al. [92] and Chakma et al. [93] proposed CNN-based models to classify outdoor images into different air quality categories. Ma et al. [94] introduced a dual-concurrent CNN model that analyzes both the original image and its transmission map for $PM_{2.5}$ estimation. Wang

et al. [95] further developed a Double-Channel CNN model that divides images into sky and building components for more accurate air quality assessment. Liu et al. [96] employed a combination of CNN and Long Short-Term Memory (LSTM) networks to analyze meteorological data and predict $PM_{2.5}$ concentrations. Li et al. [97] proposed a joint framework that integrates neural networks with Support Vector Regression (SVR) to estimate $PM_{2.5}$ levels, while Song et al. [98] presented a joint ResNet-LSTM model that combines sequential images and meteorological parameters for the same task. Chakma et al. [78] introduced a deep convolutional neural network(CNN) model to classify natural images into different categories based on their $PM_{2.5}$ concentrations. They proposed a dataset consisting of 591 daytime images from Beijing with corresponding $PM_{2.5}$ concentrations for training the model.

However, despite the advancements made by these models, there are several limitations. Many of these models fail to account for environmental conditions such as temperature, rain, haze, and other factors that impact visibility and air quality. Additionally, except for Song et al. [98], most of the existing models primarily focus on daytime images, where lighting conditions are more favorable for image-based analysis. Nighttime images, on the other hand, pose a significant challenge due to low-light conditions that can severely affect the accuracy of computer vision models.

To address these challenges, this paper proposes a novel solution for estimating $PM_{2.5}$ and PM_{10} concentrations from nighttime images by tackling two key subproblems. First, we introduce a high-low frequency module designed to extract both high and low-frequency features from the images, improving the model's ability to capture essential details even in low-light conditions. Second, we propose a Depth-Structural module that focuses on extracting structural and depth information from the images, which is crucial for accurately estimating particulate matter concentration, especially when considering the impact of buildings and other structures on the visual data. Advancement in deep learning has provided an alternative to conventional methods, leading to the development of many depth estimation methods [99–102]. The past works in the depth estimation field have been categorized into two classes.

Supervised Monocular Depth Estimation

Supervised monocular depth estimation techniques typically rely on pixel-to-pixel depth prediction to produce a depth map. Several related works [99, 103] have utilized convolutional neural networks (CNNs) to compute the learnable parameters for generating depth maps by minimizing a loss function that compares the predicted output to the ground truth. For example, in [104], a spacing-increasing discretization

technique was proposed to estimate the depth map from a color image. Additionally, Adabins [105] introduced a strategy where the depth range is divided into different bins, and a transformer is used to predict the depth profile. Other approaches, such as those in [106,107], have explored multi-model strategies for learning depth estimation.

Self-Supervised Monocular Depth Estimation

Self-supervised methods shift monocular depth estimation into an image reconstruction task, using monocular video sequences to supervise and train neural networks. Garg et al. [108] were among the first to generate a depth profile from a single color image within this framework. In 3Net [109], a trinocular stereo approach was introduced to address occlusion issues present between stereo image pairs. Monodepth [110] proposed a new loss function to enforce consistency between disparities across left and right stereo images, while Monodepth2 [111] introduced auto-masking and re-projection losses to mitigate challenges with moving objects and occlusions. Depth Hints [112] used the Semi-Global Matching (SGM) approach to produce depth priors, helping the model learn more accurate weights. EPCDepth [113] focused on horizontal image positions using data grafting techniques to improve depth map accuracy. In [7], Zhou proposed an efficient neural network model aimed at achieving monocular depth estimation.

Furthermore, we introduce a highly diverse dataset that covers multiple locations near air quality monitoring stations, encompassing a wide range of structural content and varying image depths. This dataset spans different weather conditions throughout the year, with $PM_{2.5}$ and PM_{10} concentrations varying from 3.8 to 736.4 $\mu g/m^3$ and 15.3 to 982.7 $\mu g/m^3$, respectively. The diversity of weather conditions in the dataset equips the proposed model to handle the effects of different environmental factors, such as haze, temperature, and humidity, on air quality estimation, ultimately improving the model's robustness and accuracy.

4.3 Dataset

Acquisition of the dataset has already been described at length in Section 3.3.3.

Any external activity that can lead to a change in the concentration is expected to be captured in the images, making the dataset robust. Images acquired from 6:00 PM to 6:00 AM have been considered as nighttime images. Each image is annotated with $PM_{2.5}$ and PM_{10} values from the measuring station closest to the date and time of the image taken. It may be noted that both the camera and the air monitoring station

remain well within the atmospheric boundary layer at all seasons and are hence influenced by the same surface processes. Image acquisition over different seasons makes the model robust and able to handle variations in atmospheric factors.

4.4 The Proposed Method

The framework of our proposed model is shown in Fig. 4.1. The Depth Structural (DS) module known as REAL-Net model considers the presence of multiple structures at varied horizontal depths while generating a depth map. The generated map is passed through the Residual-Attention (RA) block with a three-layer residual block added to its input with necessary up-sampling or down-sampling to match it with the feature dimensions generated by the High-Low Frequency (HLF) module. The HLF module contains five blocks with different scales that help our model extract the high and low-frequency features, concatenated with the depth-attentive features. Finally, the concatenated feature map is passed through a two-layer feed-forward network (FFN) with the ReLU [114] activation function sandwiched between them and generating a single value. The final estimate is the average of the five values generated from different scaled features.

4.4.1 High-Low Frequency Block

The HLF block shown in Fig. 4.2 inside the yellow boundary is responsible for extracting the high and low-frequency features of the image. A higher particulate concentration leads to more scattering of light, which reduces the intensity of light that reaches the surface and increases the haze or dimming of the atmosphere, introducing more low-frequency components. With fewer particles, the scattering is less pronounced, and more light reaches the surface, leading to clearer skies and higher visibility, introducing more high-frequency components. The input features are initially passed through a three-layered residual block. The low-frequency feature map covers the smooth areas of the image, including both dark regions as well as those containing the impact of light effects such as floodlight, glow, etc. The high-frequency feature map captures edges, texture, and most importantly, the presence of noise introduced due to particulate matter. To generate the high-frequency feature map, we first use two dilated convolutions [115] with filter sizes 3×3 and 5×5 with dilation rates 1 and 2 respectively, to capture features in different receptive fields, yielding the same size output denoted by f_{d1} and f_{d2}, respectively. We then subtract the two outputs and pass the results through a non-linear activation function, tanh, as stated in Eq. (4.4.8),

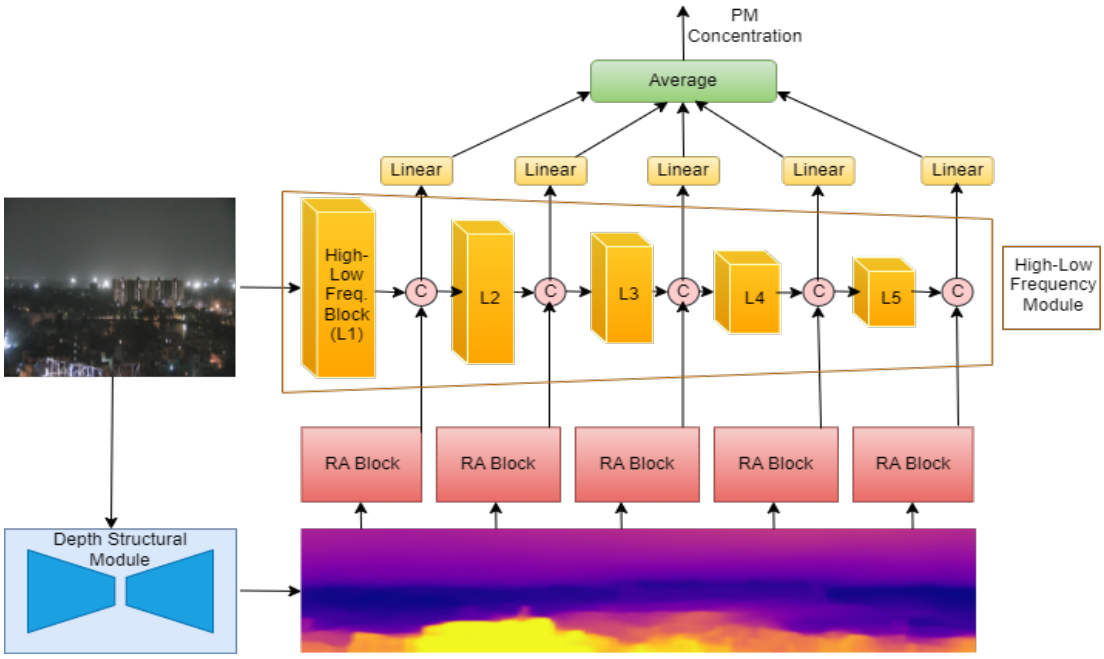


Figure 4.1: The proposed Dark-AIRNET model processes a nighttime image through five layers of Residual-Attention Blocks (RA Blocks), incorporating High-Low Frequency and Depth Structural Modules. The outputs of these modules are concatenated and passed through a linear layer to estimate the concentrations of $PM_{2.5}$ and PM_{10}. The final estimate is generated by averaging the outputs from different scales.

indicating the pixel-wise relative contrast information. The tanh function has been chosen due to its better performance over the sigmoid function in multi-layer neural networks.

$$C_a = tanh(f_{d1} - f_{d2}) \tag{4.4.1}$$

This contrast information C_a is then multiplied with the input feature map x_h to capture the high-frequency components h_f as displayed in Eq. (4.4.2), resulting in the high-frequency feature map:

$$h_f = x_h.C_a \tag{4.4.2}$$

For the low-frequency features, l_f is computed using the Eq. (4.4.3):

$$l_f = x_h.[1 - C_a] \tag{4.4.3}$$

To optimize GPU memory usage and computational complexity, the high and low-frequency feature maps are compressed via a max-pooling operation, producing h'_f and l'_f respectively. The compressed feature maps are then passed through separate convolutions, resulting in high-frequency convolutions f_{h1} and f_{h2} and low-frequency convolutions f_{l1} and f_{l2}. For each frequency, the outputs of the convolutions are multiplied using the Kronecker product, yielding z_1 for high frequency and z_2 for low frequency as shown in Eq. [4.4.4-4.4.5].

$$z_1 = f_{h1}(h_f)^T \times f_{h2}(h_f) \tag{4.4.4}$$

$$z_2 = f_{l1}(l_f)^T \times f_{l2}(l_f) \tag{4.4.5}$$

Finally, the high and low-frequency features are concatenated to form the frequency-aware feature map F as shown in Eq. (4.4.6).

Formally, given $h_f, l_f \in \mathbb{R}^{H' \times W' \times C'}$, the non-local pixel-to-pixel encoding can be summarised as

$$F = Concat(z_1, z_2) \tag{4.4.6}$$

where z_1 and z_2 are the pixel affinity maps considering the relations of each pixel to all other pixels. Finally, we obtain the frequency-aware feature map F generated by concatenating the high and low feature responses, facilitating the learning process.

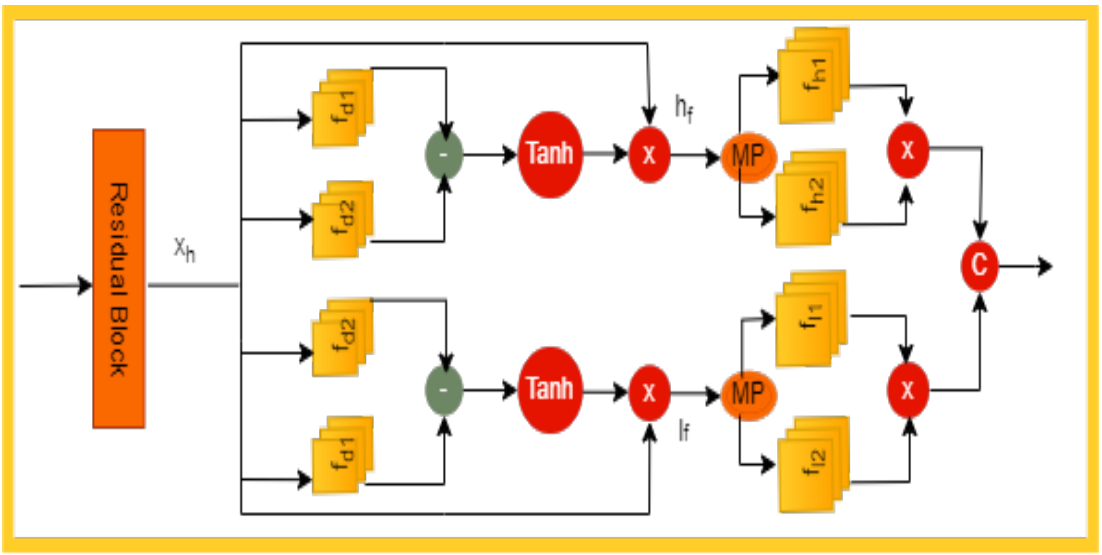


Figure 4.2: Architecture of High-Low Frequency block extracting the frequency-aware features.

4.4.2 Depth-Structural Module (REAL-Net)

Buildings at far distances exhibit low visibility, whereas close structures exhibit higher clarity. The depth module generates a depth profile from a color image. An outdoor scene may contain structures lying at varied depths and heights. Low particulate concentration leads to clear visibility, whereas high concentration results in low visibility. With the non-availability of depth information, the model gets confused with low visibility due to far distance, and high concentration. Thus, the depth of information plays a vital role in accurate estimations of the concentration, justifying the use of depth information. Further, the monitoring station also has a spatial limitation in concentration measurement, making it necessary to place multiple stations to cover a large region.

The problem of generating the depth profile [99] from a color image can be formulated as a non-linear mapping function $\boldsymbol{f}$ that maps an input image $\boldsymbol{I}$ to a real-valued depth map $\boldsymbol{D}$ as described in Eq. (4.4.7). The goal is to learn the mapping function $\boldsymbol{f}$ from a set of training data $\boldsymbol{T} = \{I_i, D_i\}_{i=1}^{M}$ with M being the total number of image pairs in the dataset T. The function $\boldsymbol{f}$ learns the weights that allow it to generate the depth profile D_i from the input image I_i.

$$\boldsymbol{f} : \boldsymbol{I} \rightarrow \boldsymbol{D} \tag{4.4.7}$$

The task of producing a depth profile for an outdoor image, as shown in Fig. 4.3, has been proposed using a neural network-based approach. An encoder-decoder model is used to estimate the depth profile of the input color image. The depth map estimation is achieved by leveraging the features from the final encoded layer, denoted as E_L. This layer is derived from a set of intermediate multi-scale features $\{E_l\}_{l=1}^{L-1}$. Specifically, $\mathbf{E} = \{\mathbf{E_l}\}_{l=1}^{L}$, with each $\mathbf{E_l} = \{e_l^i\}_{i=1}^{N}$ and $e_l^i \in \mathbb{R}^{Cin}$ where L is set to 5 and Cin represents the input number of channels. Attention maps A_l are associated with each intermediate encoded layer E_l, which help selectively regulate the information passed from the intermediate layers relevant for depth estimation. Here, $\mathbf{A} = \{\mathbf{A_l}\}_{l=1}^{L-1}$, with each $\mathbf{A_l} = \{a_l^i\}_{i=1}^{N}$ and $a_l^i \in \mathbb{R}^{Cin}$. We enforce a regional constraint that ensures neighboring pixel attention variables are related to enhance the extraction of region-specific information. Finally, we define an energy function that integrates the intermediate encoded layers with their attention maps, as described in Eq. (4.4.8).

$$\Gamma(E, A) = \phi(E, A) + \psi(A) \tag{4.4.8}$$

The first term $\phi(E, A)$ establishes the relationship between the intermediate and final

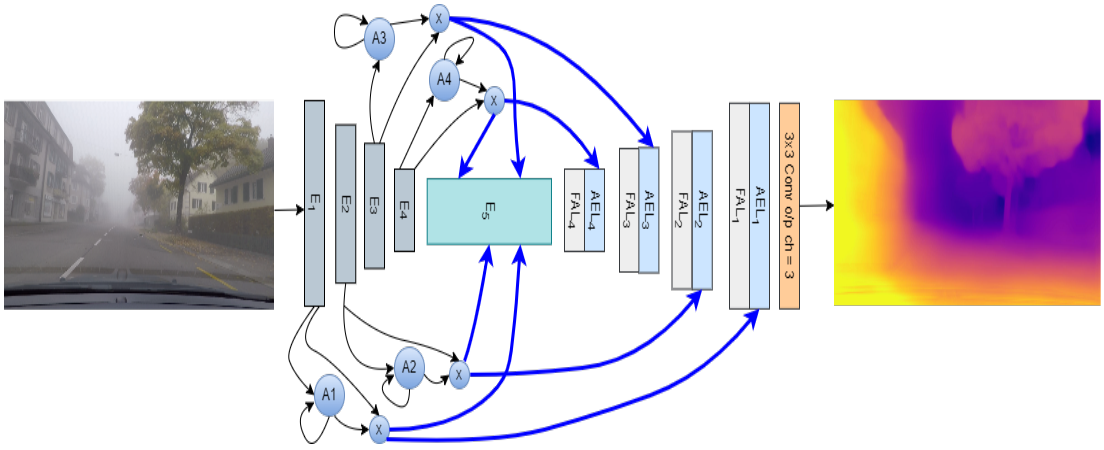


Figure 4.3: Architecture of Monochromatic Depth Estimation module extracting the depth-based regional information from the encoder. The model utilizes the extracted attention-encoded features and spatial-feature value attention layer in the decoder to produce the estimated depth map. The deep blue arrows represent the flow of attention-guided information.

encoded layers, incorporating the attention maps. This regulates the features that are essential for generating the depth profile. It is defined as:

$$\phi(E, A) = \sum_{l=1}^{L-1} \xi(\mathbf{A}_l, \mathbf{E}_l) \tag{4.4.9}$$

$$\xi(\mathbf{A}_l, \mathbf{E}_l) = \sum_{i,j} a_l^i \beta_{i,j}^l \mathbf{e}_l^j \tag{4.4.10}$$

Here a_l^i represents the attention variables and $\beta_{i,j}^l \in \mathbb{R}^{C_l \times C_l}$ where C_l denotes the channel dimension of E_l. The final term $\psi(\mathbf{A})$ introduces a constraint to the energy function to capture regional homogeneity, ensuring the relationships among the neighboring pixels in the attention map are properly captured. The term is defined as follows:

$$\psi(A) = \sum_{l=1}^{L-1} \sum_{i,j} \kappa_{i,j}^l a_l^i a_l^j \tag{4.4.11}$$

Here $\kappa_{i,j}^l$ is a trainable parameter.

Attention Approximation

Given the energy function defined in (4.4.8), finding its optimized solution is difficult because its distribution is unknown. The Mean Approximation strategy is used to address this issue where a new distribution q is estimated to minimize the KL di-

vergence between the distribution of $\mathbf{\Gamma}$ and q.

Encoding Strategy

Mean-field approximation technique derives the updates for the final encoded layer and the attention variables of each intermediate encoded layer. $\mathbf{E}_q$ referred to as the expectation over the distribution q is obtained as:

$$\begin{aligned} \log\{q(e_L^i)\} &\propto \sum_{l\neq L} \mathbf{E}_{q(a_l^i)}\{a_l^i\} \sum_j \mathbf{E}_{q(e_l^j)}\{\xi(a_l^i, e_l^j)\} \\ \log\{q(a_l^i)\} &\propto a_i^l \sum_j \mathbf{E}_{q(e_L^j)}\{\xi(a_l^i, e_L^j)\} + \sum_l \sum_j \mathbf{E}_{q(a_l^j)}\{\kappa_{i,j}^l a_l^i a_l^j\} \end{aligned} \tag{4.4.12}$$

Considering the distributions $q(e_L^i)$ and $q(a_l^i)$, $\mathbf{E}_{q(a_l^i)}\{a_l^i\}$, $\mathbf{E}_{q(e_l^i)}\{e_l^i\}$ and $\mathbf{E}_{q(e_L^i)}\{e_L^i\}$ are referred to by $\mathbf{a_l^i}$, $\mathbf{e_l^i}$ and $\mathbf{e_L^i}$ respectively.
Considering the potential defined in (5.3.3) we write:

$$\mathbf{e_L^i} = \sum_{l\neq L} \sum_{i,j} \mathbf{a_l^i} \beta_{i,j}^l \mathbf{e}_l^j \tag{4.4.13}$$

By considering the potential defined in equations (4.4.11) and (4.4.13), the attention variables are computed using the sigmoid function, as shown in equation (4.4.14). This ensures that the variables are normalized:

$$\mathbf{a_l^i} = \sigma(\sum_j \beta_{i,j}^l \mathbf{e_L^j} + \sum_j \kappa_{i,j}^l \mathbf{a_l^j}) \tag{4.4.14}$$

Thus, the model computes the attention variables by applying a sigmoid function to the extracted features while considering the relationships among different a_l^i to enforce the regional constraint. This approach enables the simultaneous learning of both the feature and attention vectors, along with the parameters κ and β, which help maintain the constraints. Down-sampling operations are applied to the intermediate encoding, resulting in the Attention Encoded Layer (AEL), which is ultimately used to generate the final encoded layer, E_5.

Decoding Strategy

The decoder performs the up-scaling process that increases the spatial dimensions to their original shape through four decoder blocks. Each block includes an Attention Encoded Layer (AEL) and a Feature Attention Layer (FAL). The AEL holds the encoded feature maps generated by the product of the encoded layer E_l and A_l preserving both fine and semantic details of the image.

In the conventional convolutional layer [116], a kernel is applied across the entire feature map, allowing the kernel weights to be shared globally. Given $I = (i_1, i_2, \cdots, i_N)$, $i_v \in \mathbb{R}^{c_{in}}$ represents a decoded feature vector with N pixels and c_{in} channels, the spatial convolution with filter weight $W \in \mathbb{R}^{c_{out} \times c_{in} \times p \times p}$ is computed as shown in (4.4.15). The weights are shared spatially but not synchronized with the feature values.

$$i_u^{'} = \sum_{v \in N(u)} W[s_u - s_v] i_v + b \tag{4.4.15}$$

Here, $\mathcal{N}(u)$ refers to pixels in the neighborhood of u with patch size $p \times p$, $s_u = (x_u, y_u)$ are pixel coordinates, $b \in \mathbb{R}^{c_{out}}$ refers to the biases, $(s_u - s_v)$ points to the spatial indexing of an array with 2D spatial offsets, $i_u^{'} \in \mathbb{R}^{c_{out}}$ is the output of the convolution operation for each pixel u.

To address this limitation the decoding strategy modifies the weight by multiplying it with a factor that incorporates the feature values of the corresponding pixel coordinates, as described in (4.4.16). The Feature Attention Layer(FAL) uses a local weight factor computed based on the features f_u and f_v where $v \in \mathcal{N}(u)$, represents the neighboring pixels of pixel u.

$$i_u^{'} = \sum_{v \in \mathcal{N}(u)} \chi(f_u, f_v) W[s_u - s_v] i_v + b \tag{4.4.16}$$

$$\chi(f_u, f_v) = e^{-\frac{1}{2}(f_u - f_v)^T (f_u - f_v)} \tag{4.4.17}$$

where χ is a fixed-parametric function in the form of a Gaussian curve, applied to feature values f_u, f_v at indices u and v respectively, but belonging to $\in \mathbb{R}^{c_{in}}$. This filtering is performed on each c_{out} 2D matrix. This operation is referred to as the Feature Attention layer (FAL). The FAL considers the spatial and feature variations among the neighboring pixels. This results in higher weights for features with similar values, aligning to maintain regional constraints in the output image.
Each decoder block l consists of FAL applied to its previous layer, followed by the concatenation of the AEL of encoded layer l as shown in Fig. 4.3. Finally, a 3×3 convolution operation generates the depth map.

This pre-trained model has been specifically trained on a nighttime dataset to generate accurate depth maps corresponding to nighttime images as shown in Fig. 4.4. To ensure the consistency and quality of the generated depth map, we compute the depth consistency loss using two key components: the L_1 norm and the Structural Similarity

Index Measure (SSIM).

The L_1 norm measures the pixel-wise absolute difference between the predicted and ground truth depth maps, providing a simple yet effective way to quantify the error. The SSIM, on the other hand, is used to measure the structural similarity between the generated and ground-truth depth maps. By combining these two metrics, the model can ensure both pixel-level accuracy and structural integrity, which are crucial for estimating particulate matter concentrations from nighttime images. The depth consistency loss function can be represented as follows:

$$\mathbf{L} = \alpha(1 - SSIM(\mathbf{D}_{pred}, \mathbf{D}_{gt})) + (1 - \alpha)|\mathbf{D}_{pred} - \mathbf{D}_{gt}| \tag{4.4.18}$$

Here, $\mathbf{D}_{pred}$ and $\mathbf{D}_{gt}$ are the predicted and ground truth depth maps of the input pair image, with α being the hyperparameter set at 0.75.

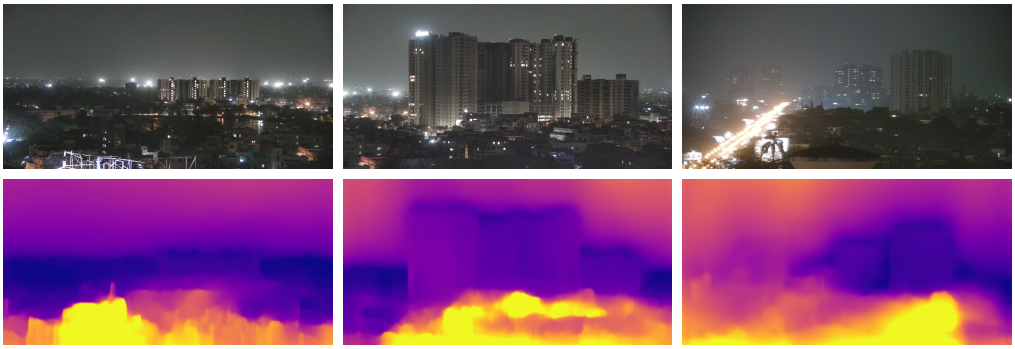


Figure 4.4: Depth map presentation of night images with the top row indicating color images and the bottom row with their corresponding depth maps.

4.4.3 Residual Attention Block

To effectively handle multi-scale features and preserve spatial details while avoiding the loss of important dependencies on high-level features, we have designed a Residual Attention Block (RAB). This block facilitates the learning of high-level features and, through skip connections, mitigates the training difficulties commonly associated with deep networks.

The architecture of the Residual Attention Block (as shown in Fig. 4.5) consists of two key components arranged in a series format: the Pixel Attention Module and the Channel Attention Module [117]. These modules are designed to enhance feature

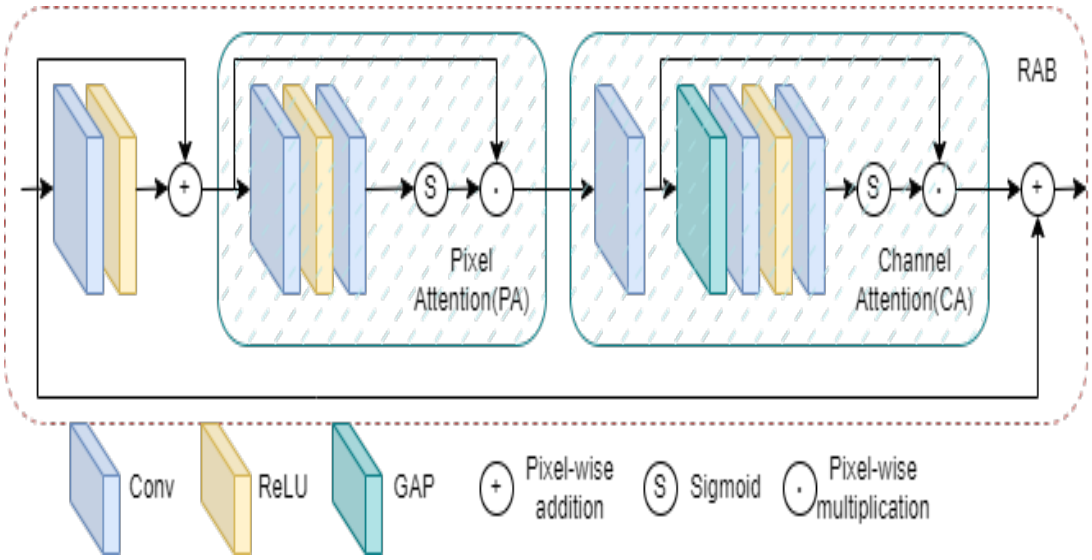


Figure 4.5: Architecture of Residual Attention Block(RAB).

representations while maintaining critical spatial details.

A key aspect of the design is that no up-sampling or down-sampling is performed within the attention block itself. This ensures that spatial details are preserved throughout the learning process. Convolution layers within the attention block utilize a kernel size of 3×3 with padding set to 1, to maintain the input's spatial resolution and to avoid introducing additional distortions or information loss.

However, to meet the size constraints during the concatenation of features from the high-low frequency aware module with the attention-processed features, down-sampling is carried out at the pre-attention mechanism stage. This ensures that the feature maps are appropriately sized before they are passed into the attention modules, while still preserving the integrity of the learned features.

By leveraging the residual nature of the attention block and using skip connections [118], the model can efficiently learn deep features without the degradation typically observed in deeper networks. This approach contributes significantly to maintaining both fine-grained spatial information and high-level feature dependencies, leading to a more robust model for estimating particulate matter concentration from nighttime images.

4.4.4 Multi-scale Fusion

The multi-scale outputs, resulting from the concatenation of the high-low frequency aware features and the attention-based depth features, are then passed through the corresponding Linear layer, which is composed of a 64-dimensional 2-layer multi-level perceptron. This allows the model to make predictions for the particulate matter concentration at different scales. Rather than directly using the output from just the last scale feature module, we propose a dynamic fusion mechanism that integrates the predictions from all the scales, taking advantage of the features at each scale.

The final prediction of the concentration is obtained by calculating the weighted average of the five outputs from the different scales. The weights for the averaging process are learnable, allowing the model to adaptively focus on the most relevant scales during training, ensuring that no single scale's features are overused or under-utilized. This mechanism helps the model make better use of the diverse information contained in the multi-scale features.

Mathematically, this dynamic fusion mechanism can be expressed as in:

$$F_{Final} = \sum_{i=1}^{5} \omega_i F_i \tag{4.4.19}$$

where F_1, F_2, F_3, F_4, and F_5 are the feature outputs from different scales and ω_1, ω_2, ω_3, ω_4, and ω_5 are the learnable weights associated to each scale's importance.

This fusion strategy enables the model to combine features from multiple scales effectively, ensuring that the final concentration estimation benefits from the strengths of all scales rather than being dominated by any single one.

4.5 Experiment and Evaluation

This section focuses on the experimental settings used in developing the method. This includes the performance evaluation metrics, model implementation details, and the ablation study.

4.5.1 Training Set

Thorough test and validation of the proposed depth estimation model have been achieved with the help of several benchmark datasets, including KITTI [4], Cityscape [8], and Foggy Zurich [119]. For training, validation, and testing, the approach mentioned in [99], utilizes 22,000 images from the KITTI dataset for training and 700 im-

ages for testing. An appropriate training dataset is a major requirement for the particulate estimation model's training. We have framed certain rules to select images from the large dataset prepared for DarkAirnet's development and training. The criteria followed to prepare the training dataset are as follows:

1. Images from different views such that images contain structures near, medium, and far.

2. Images were recorded during different seasons of the year. These inputs have been selected for each month, covering summer, rainy, autumn, winter, and spring conditions.

3. Images recorded at dates and times with varied particulate concentration levels.

4. Images with different shapes of structures with varied lengths.

5. Images during certain rare events when the particulate concentration is extremely high such as fire bursting during Diwali or other festivals, rainfall during winter, etc.

Following the above criteria, a curated dataset of 3790 images was used for training the particulate estimation model, which was then tested on 48,200 images as the test set. The training set created following this procedure consists of all different kinds of scenarios that play a vital role in making the model accustomed to varied scenes. The number 3790 was chosen to optimize the availability of scanty computational resources, chiefly GPU time. Once the model was trained with the training dataset, we computed the concentration values for all those images that had not been used for training, specifically 48,200 images, discarding images for which reference PM values were not available.

4.5.2 Loss function

In our approach, we use the pre-trained depth estimation model alongside the High-Low Frequency module to train on our proposed dataset. This training allows the model to more accurately estimate the concentration of particulate matter ($PM_{2.5}$ and PM_{10}) based on nighttime images, improving the estimation accuracy by leveraging both depth and frequency-aware features.

The concentration of particulate matter ($PM_{2.5}$ and PM_{10}) varies widely, ranging from very low to extremely high levels, often reaching four-digit concentrations in

many cities and towns in India. The distribution is highly skewed, with the highest concentrations typically occurring during the winter months. Most of the time, the concentration remains within a range of [0-200] $\mu g/m^3$ while extreme values are rare events. It is important to note that incorrect estimates tend to magnify the loss due to the squaring term in Mean Squared Error (MSE). However, large errors from these rare high-concentration events may disproportionately impact the model's performance due to the equal weightage given to both normal and extreme values, as in Mean Absolute Error (MAE). To address this issue, we opt to use the Huber loss function for model optimization, as it reduces the influence of outliers while maintaining robustness in the estimation process.

$$L(p^e, p^{gt}) = \begin{cases} \frac{(p^{gt}-p^e)^2}{2}, & \text{if } |p^{gt} - p^e| < \delta \\ \delta|p^{gt} - p^e| - \frac{\delta^2}{2}, & \text{otherwise} \end{cases}$$

where p^e, p^{gt} indicates the estimated and the actual concentration of particulate matter in the air of the given region, and the δ parameter is set to 0.5 throughout the process. Thus, if the error between the estimated and actual is < 0.5, MSE is used elsewhere in MAE.

4.5.3 Implementation details for REAL-Net

We implemented our proposed method on the Pytorch platform, with the input image size fixed at 640×480 during training. The model was optimized using the Adam optimizer, with an initial learning rate of 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$. Training was performed on an NVIDIA GeForce RTX 2060-8GB GPU, with a batch size of 16 and 100 epochs.

4.5.4 Comparison methods for depth estimation model

The section presents a comparison of the proposed method with recent works, as illustrated in Fig. 4.6 and 4.7. Quantitative evaluation results are provided in Tables 4.2, 4.3, and 4.4 for the KITTI, Cityscape, and Foggy Zurich datasets, respectively, with the results obtained by our method highlighted in bold. We evaluate performance using three metrics: Accuracy (δ), Mean Relative Error (REL), and Root Mean Squared Error (RMS), as described below. Given the test data, where $\hat{d}_i$ and d_i denote the estimated and ground-truth depth values for pixel i:

1. Accuracy (δ): The percentage of pixels where $\max(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}) < t$ with $t \in [1.25, 1.25^2]$.

Method	*REL*	*RMS*	$\delta < 1.25$	$\delta < 1.25^2$
(a)	0.217	5.917	0.611	0.632
(b)	0.197	4.988	0.683	0.695
(c)	0.152	4.712	0.750	0.797

Table 4.1: Ablation Study on the KITTI dataset.

2. Mean Relative Error (REL): $\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{d}_i - d_i|}{d_i}$.

3. Root Mean Squared Error (RMS): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{d}_i - d_i)^2}$

4.5.5 Ablation Study on REAL-Net

To demonstrate the effectiveness of the proposed method, an ablation study was performed on the KITTI dataset, with the quantitative results presented in Table 4.1. The study evaluates the following three cases: (a) excluding the attention mapping in the encoder and the local mapping in the decoder, (b) excluding the attention mapping in the encoder, and (c) excluding the local mapping in the decoder.

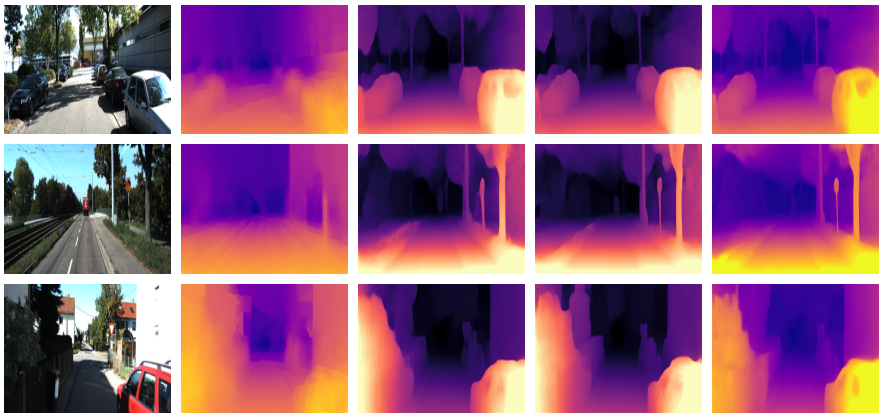


Figure 4.6: Visual comparison of depth estimation on images from the KITTI [4] dataset. The first to the fifth columns correspond to the color image and the results obtained from [5], [6], [7], and our proposed method, respectively.

4.5.6 Evaluation Metrics

The performance of the model has been evaluated using Symmetric Mean Absolute Percentage Error (SMAPE), R-squared (R^2), and Pearson Correlation(ρ) between the estimated and actual value of $PM_{2.5}$ and PM_{10}. SMAPE measures the accuracy of the

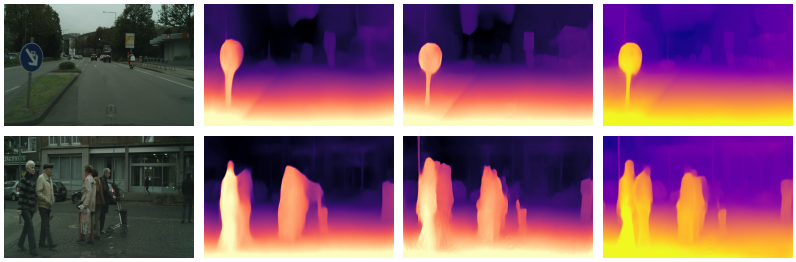


Figure 4.7: The visual comparison of depth estimation on images from the Cityscapes [8] dataset is shown. The first column presents the color image, followed by the depth results obtained from [6], [7], and our proposed method in the second, third, and fourth columns, respectively.

Method	*REL*	*RMS*	$\delta < 1.25$	$\delta < 1.25^2$
Xu et al.[5]	0.143	4.691	0.872	0.887
Monodepth2[6]	0.115	4.865	0.877	0.959
MonoVIT[7]	0.102	4.483	0.896	0.965
Proposed	**0.097**	**4.131**	**0.901**	**0.971**

Table 4.2: Quantitative visual quality performance on KITTI dataset

estimated results based on the relative error. R^2 measures the best-fit line produced across the estimated concentration values. ρ establishes the correlation between the estimated and the actual value concentration values. Lower SMAPE, higher R^2, and higher ρ values refer to a highly efficient model. The experimental results conclusively show that Dark-AirNet outperforms the other networks.

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|x_i - y_i|}{(|x_i| + |y_i|)/2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Method	*REL*	*RMS*	$\delta < 1.25$	$\delta < 1.25^2$
Xu et al.[5]	0.158	4.013	0.755	0.793
Monodepth2[6]	0.127	3.502	0.811	0.874
MonoVIT[7]	0.104	3.195	0.836	0.891
Proposed	**0.098**	**2.787**	**0.910**	**0.933**

Table 4.3: Quantitative visual quality performance on Cityscapes dataset.

Method	*REL*	*RMS*	$\delta < 1.25$	$\delta < 1.25^2$
Xu et al.[5]	0.147	1.980	0.891	0.903
Monodepth2[6]	0.131	1.964	0.916	0.928
DIFFNet[7]	0.124	1.955	0.922	0.931
Proposed	**0.112**	**1.936**	**0.936**	**0.952**

Table 4.4: Quantitative visual quality performance on Foggy Zurich dataset.

Method	SMAPE		R-squared		Pearson Coeff.	
	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}
Dark-AirNet	**0.172**	**0.168**	**0.861**	**0.868**	**0.910**	**0.921**
HLF Module	0.367	0.351	0.716	0.722	0.846	0.849
DS Module	0.255	0.281	0.771	0.789	0.866	0.871

Table 4.5: Ablation Study results of the network. The best result for each metric has been displayed in bold font.

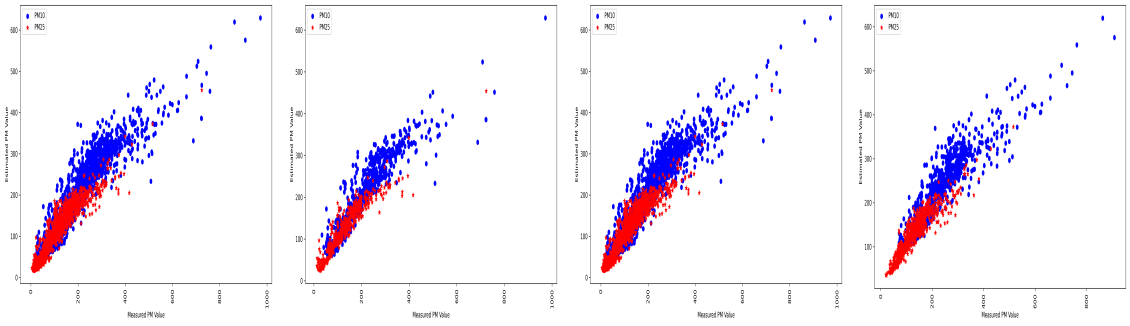
4.5.7 Observed Patterns in Particulate Matter Concentration

In Table 4.6, six categories of the concentration of particulate matter $PM_{2.5}$ and PM_{10}, as standardized by the Indian regulatory authority, have been shown. The density of particulate matter behaves differently under different seasons, exhibiting high seasonal variation. The variation is not just limited to seasons but also throughout the night. Variations in temperature, relative humidity, wind direction, and wind velocity play a vital role in the varying concentrations of particulate matter over the season.

The variations in PM as recorded at the air monitoring station and estimated by the proposed model are now examined throughout the year and separately during the winter, monsoon, and autumn seasons. Scatter plots of measured and estimated values, shown in figure 4.8, illustrate the relationship between the two.

AQI category (range)	$PM_{2.5}$	PM_{10}
Good (0–50)	0–30	0–50
Satisfactory (51–100)	31–60	51–100
Moderate (101–200)	61–90	101–250
Poor (201–300)	91–120	251–350
Severe (301–400)	121–250	351–430
Hazardous (401–500)	250+	430+

Table 4.6: AQI category, pollutants, and health breakpoints in $\mu g/m_3$



a b c d

Figure 4.8: Scatter plots of Measured vs Estimated PM in order - (a) Annual, (b) Autumn, (c) Rainy, and (d) Winter respectively.

4.5.8 Annual Observation

Fig. 4.9 and 4.10 illustrate the trend followed by the particulate matter $PM_{2.5}$ and PM_{10} follows respectively. The yellow line refers to the data measured by the air quality monitoring station, whereas the green line follows the data obtained by our Dark-AIRNET model. The average value obtained from the data taken from 6:00 PM to 6:00 AM has been considered in both cases.

Starting from July 2021, we find the air quality to remain moderate till the end of the rainy season. The onset of autumn, followed by winter, makes the quality of air poor, with January being the coldest month, taking the quality to a severe category, with certain rare cases of hazardous level. The spring season, followed by the hot summer months, makes the air quality satisfactory as observed by the monitoring station. Our model-generated data follows the same pattern as observed in the actual data. The estimated prediction differs appreciably from the actual data only in certain extreme cases, which can be seen to occur rarely. The observed results demonstrate the cyclical nature of the particulate matter throughout the year, making winter and summer the worst and the best seasons in terms of air quality.

4.5.9 Observation during the rainy season

The rainy or monsoon season, as the term is used in India, starts in June with the onset of the southwest monsoon winds and continues till the end of August. High wind speeds and severe rainfall during the monsoon season make the particulate matter settle on the surface of the earth, leading to lower concentration levels despite having high relative humidity and lower surface temperature. We have shown the trends

observed from the data generated by the monitoring station and our proposed model. The air quality remains at a satisfactory level in the majority of cases. Fig. 4.11 and 4.12 show the expected trend observed during the majority of the time. Underestimation of the particulate concentration is mostly observed to have happened where sudden surges in this concentration have occurred.

4.5.10 Observation during the autumn season

The autumn season, or post-monsoon, starts in September with the retreat of the southwest monsoon wind till the end of October. Low-speed winds and a fall in temperature observed during the time make the concentration of particulate matter start rising slowly and steadily. We have shown the trends observed from the data generated by the monitoring station and our proposed model. The air quality remains within the poor category level in most cases. Fig. 4.13 and 4.14 show the expected trend observed during the majority of the time.

In the figures, it can be observed that during the intervening night on the 4th and 5th of November 2021, a sudden spike in the particulate concentration is observed. This was due to the burning of firecrackers on the eve of the Diwali festival, celebrated throughout the nation. The model output very effectively followed the same pattern as that computed by the monitoring station. Further, during the interval from 5th to 7th of December 2021, the sudden fall in the particulate concentration was mainly due to untimely rainfall, and the same trend is observed in the resultant output.

4.5.11 Observation during the winter season

Winter months in India mainly stretch from mid-December to mid-March, with January and February among the coldest months. A lower wind speed, low temperature, and high relative humidity trap the particulates near the surface of the earth as the air becomes heavier, thus preventing rapid movement. This leads to a sharp rise in the concentration of pollutants to severe levels. In Fig. 4.15 and 4.16, the trend indicates that the majority of the time, the concentration of particulates $PM_{2.5}$ and PM_{10} remains above the poor category.

4.5.12 Model Settings

The model has been trained using 3790 images randomly selected from the proposed dataset, such that it covers a wide range of $PM_{2.5}$ with different scenes and weather conditions. We have used high-quality, large-size color images of size 1024×1024 as

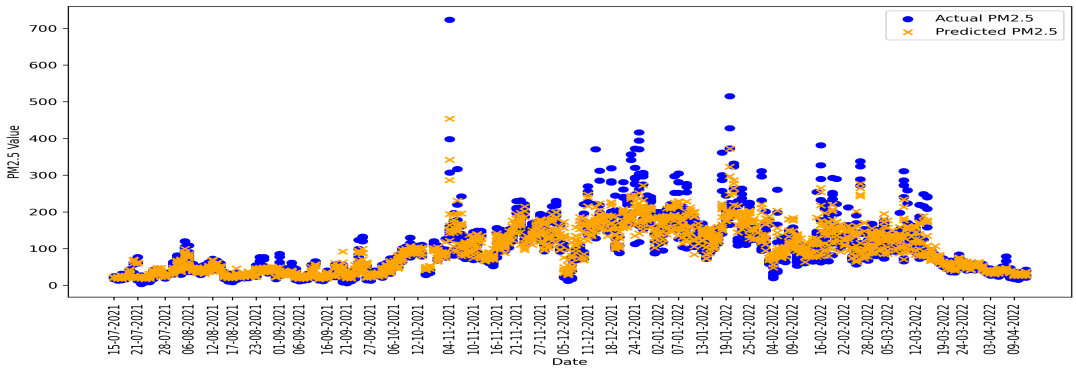


Figure 4.9: Pattern exhibited in $PM_{2.5}$ data collected throughout the year between July 2021 and April 2022

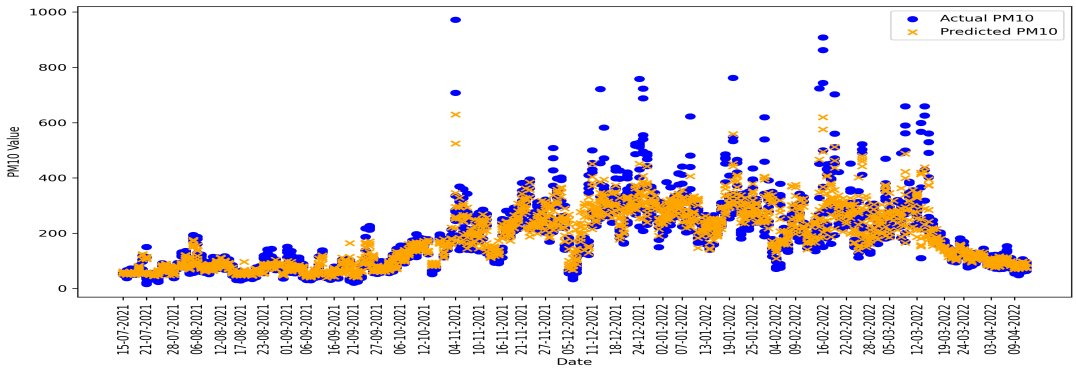


Figure 4.10: Pattern exhibited in PM_{10} data collected throughout the year between July 2021 and April 2022

input to the model and further augmented the training dataset with random rotation by an angle of 45, 90, 180, and 270. The proposed model has been optimized using Adam optimizer with the initial learning rate set to 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$ and $\delta = 0.5$. The batch size has been set to 16, and the total epoch number has been set to 500, with 207 iterations in each epoch. The GPU used for the task was the Nvidia GeForce RTX 2060 Super with 8GB RAM.

4.5.13 Results: Performance comparison and quantitative assessment

Table 4.7 shows the comparison of the performance of our model with multiple state-of-the-art CNN architectures viz Resnet-50 ([120]), Densenet-121 ([121]), Inception-V4 ([122]), Inception-ResNet-V2 ([122]), and EfficientNetB7 ([123]). The experimental results conclusively show that Dark-AIRNET outperforms the other networks.

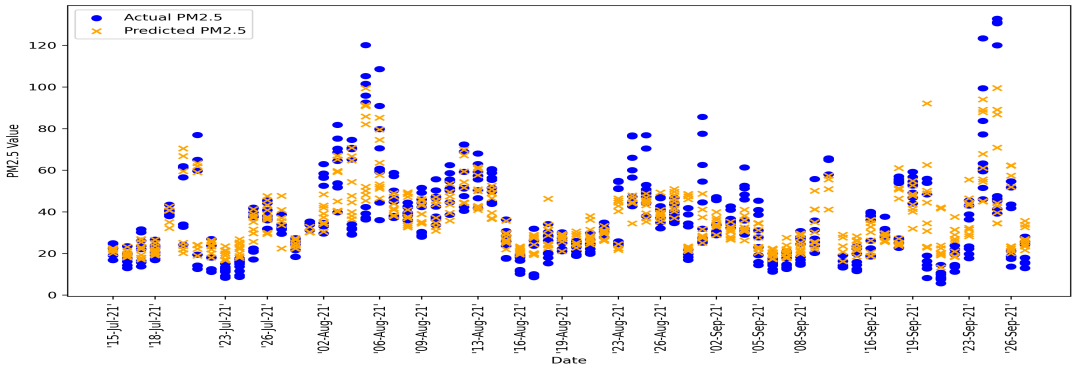


Figure 4.11: Pattern exhibited in $PM_{2.5}$ data collected during the period of July-September, 2021

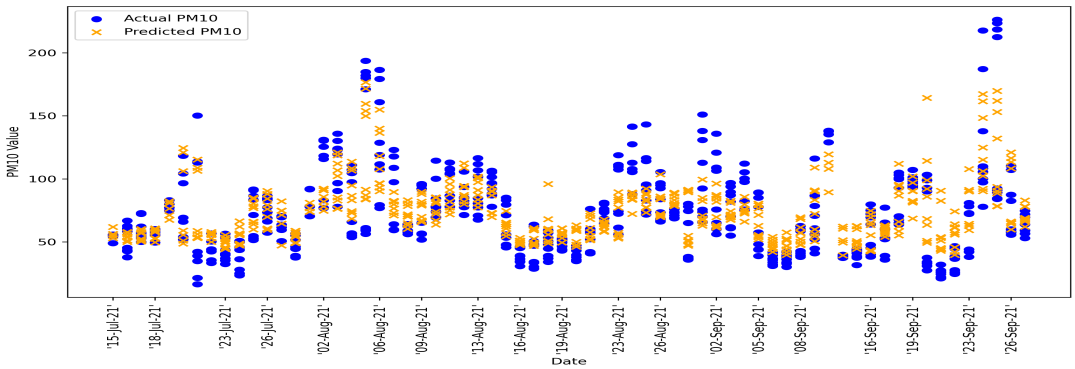


Figure 4.12: Pattern exhibited in PM_{10} data collected during the period of July-September, 2021

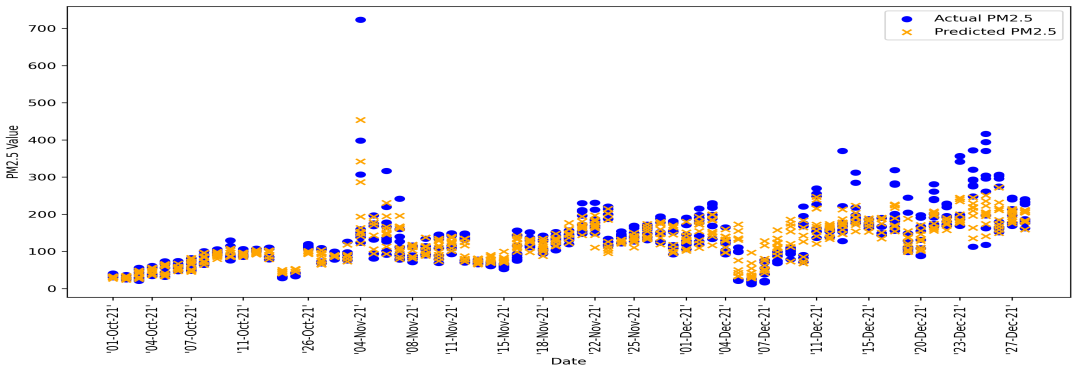


Figure 4.13: Pattern exhibited in $PM_{2.5}$ data collected during the period of October-December, 2021

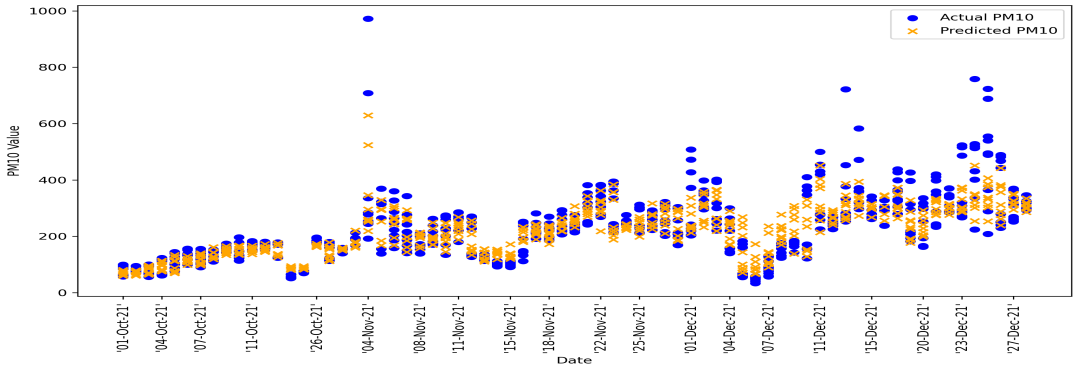


Figure 4.14: Pattern exhibited in PM_{10} data collected during the period of October-December, 2021

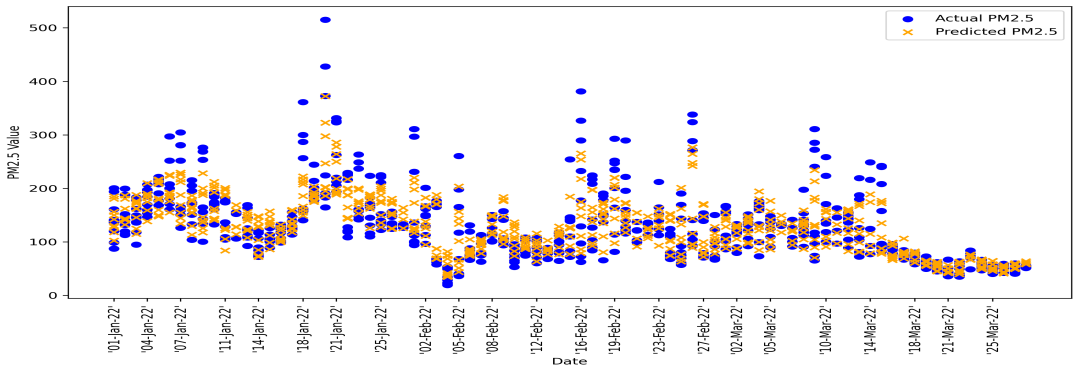


Figure 4.15: Pattern exhibited in $PM_{2.5}$ data collected during the period of January-March, 2022

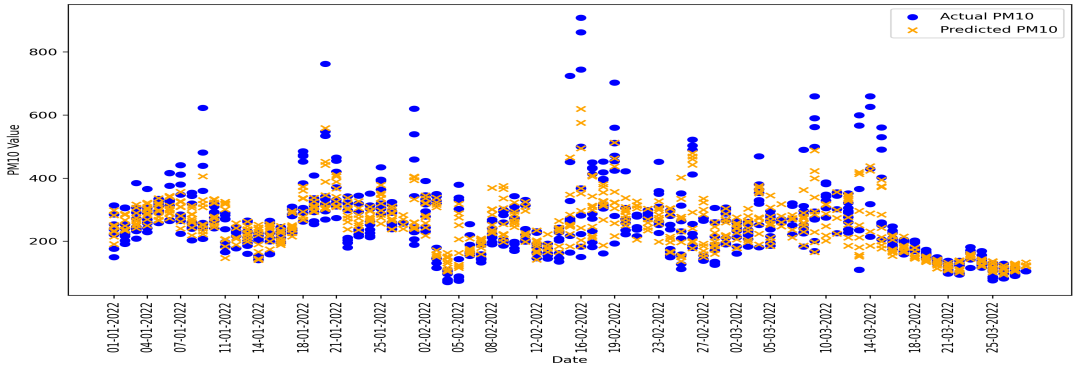


Figure 4.16: Pattern exhibited in PM_{10} data collected during the period of January-March, 2022

Method	SMAPE		R-squared		Pearson Coeff.	
	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}
Resnet-50[120]	0.246	0.251	0.817	0.823	0.883	0.886
Resnet-101[120]	0.308	0.323	0.812	0.818	0.869	0.871
Resnet-152[120]	0.335	0.333	0.768	0.776	0.853	0.857
Densenet-264[121]	0.711	0.703	0.317	0.319	0.391	0.396
Densenet-201[121]	0.702	0.698	0.329	0.331	0.406	0.403
Densenet-169[121]	0.683	0.671	0.347	0.351	0.417	0.411
Densenet-121[121]	0.681	0.670	0.351	0.357	0.429	0.432
Inception-V3[124]	0.480	0.487	0.551	0.548	0.736	0.738
Inception-V4[122]	0.485	0.492	0.555	0.562	0.745	0.749
Inc-ResNet-V1[122]	0.411	0.388	0.542	0.531	0.738	0.731
Inc-ResNet-V2[122]	0.419	0.401	0.555	0.547	0.746	0.741
EfficientNetB5[123]	0.413	0.419	0.672	0.681	0.811	0.826
EfficientNetB6[123]	0.391	0.311	0.686	0.688	0.828	0.833
EfficientNetB7[123]	0.385	0.307	0.691	0.698	0.837	0.842
Dark-AirNet	**0.172**	**0.168**	**0.861**	**0.868**	**0.910**	**0.921**

Table 4.7: Quantitative performance achieved by models for particulate estimation. The best result for each metric has been displayed in bold font.

To quantitatively assess the estimation errors made by the proposed Dark-AIRNET, the following performance measures have been computed for each of the four seasons: summer, monsoon, autumn, and winter. $\hat{\theta}_i$ represents the estimated value of the PM concentration on the ith day, while θ_i stands for the corresponding directly measured value, and N stands for the total number of data samples (i.e., sample size) in the season under consideration.

$$\text{Average Estimated Bias} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i) \tag{4.5.1}$$

$$\text{Average Estimated Relative Bias} = \frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{\theta}_i - \theta_i)}{\theta_i} \tag{4.5.2}$$

$$\text{Estimated Root Mean Squared (RMS) Error} = \sqrt{\frac{1}{N}\sum_{1=1}^{N}(\hat{\theta}_i - \theta_i)^2} \tag{4.5.3}$$

$$\text{Estimated RMS Relative Error} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right)^2} \tag{4.5.4}$$

While the first two measures highlight the systematic bias, the next two reflect the uncertainty in the model's estimated values across different seasons. Table 4.8 presents

Pollutant	Performance Measure	Summer	Monsoon	Autumn	Winter
$PM_{2.5}$	Average Estimated Bias	1.816	0.234	5.587	2.134
	Average Estimated Relative Bias	0.138	0.130	0.045	-0.039
	Estimated RMS Error	7.157	8.865	31.502	8.947
	Estimated RMS Relative Error	0.278	0.416	0.350	0.493
PM_{10}	Average Estimated Bias	-2.664	-0.907	7.815	4.097
	Average Estimated Relative Bias	0.065	0.115	0.036	0.029
	Estimated RMS Error	13.276	16.694	54.933	16.694
	Estimated RMS Relative Error	0.175	0.359	0.265	0.359

Table 4.8: Performance Measures of Dark-AIRNET

these measures for the estimation of $PM_{2.5}$ and PM_{10} across the four seasons. The average estimated bias indicates whether the model overestimates or underestimates the pollutant concentrations compared to the measured values that are considered to be the reference. A positive value reflects a tendency to overestimate the measured values, while a negative value implies underestimation. Estimates of $PM_{2.5}$ show overestimation across all seasons, with the highest value observed in autumn (mean = 5.59). The model underestimates the value of PM_{10} in summer and monsoon, in contrast to the estimates of $PM_{2.5}$, but it can be observed that the values are small. Overestimation of PM_{10} in autumn and winter, with the error having the greatest magnitude in autumn, is observed, similar to the estimates of $PM_{2.5}$.

The Average Estimated Relative Bias measures the systematic error in pollutant estimation as a proportion of the true value, providing a scale-independent model accuracy assessment. The table shows that for $PM_{2.5}$ estimation, the measure is positive (i.e., overestimation) in summer, monsoon, and autumn, but slightly negative in winter, indicating a small underestimation during that season. For PM_{10}, it is consistently positive across all seasons, suggesting a general tendency to overestimate the true values, although the magnitude of this bias is relatively small for both $PM_{2.5}$ as well as PM_{10}.

The Estimated RMS Error is a statistical measure used to quantify the average magnitude of errors in a set of estimated and measured values, i.e., how far off the predictions are compared to the measured values. Since it squares errors before averaging, larger errors have more influence on the RMSE than smaller ones. Autumn exhibits the highest error for both $PM_{2.5}$ (31.502) and PM_{10} (54.933), due to various outlier instances of fire cracking incidents that occur during the festivals such as Di-

wali. Conversely, the summer, monsoon, and winter data show moderate variability besides exhibiting a low Average Estimated Bias and Relative Bias, suggesting relatively better model performance during these periods.

The Estimated RMS Relative Error is a performance metric that quantifies the average size of the relative errors between predicted and measured values, scaled by the actual values, using a root mean square (RMS) approach. It measures how large the prediction errors are relative to the true measured values and, by squaring the larger relative errors, gets amplified, making the metric sensitive to large deviations. As observed in Table 4.8, the relative error for PM_{10} is less compared to $PM_{2.5}$ for all seasons. Furthermore, a sharp deterioration in visual information during monsoon and winter leads to a moderate relative error with good relative accuracy in summer and autumn for both $PM_{2.5}$ and PM_{10}.

A notable observation about average estimated bias for summer and monsoon data is that though the biases are opposite in sign for $PM_{2.5}$ and PM_{10}, the magnitude-wise difference is within reasonable limits. Similarly, though the average estimated relative bias is negative but the magnitude-wise difference is negligible.

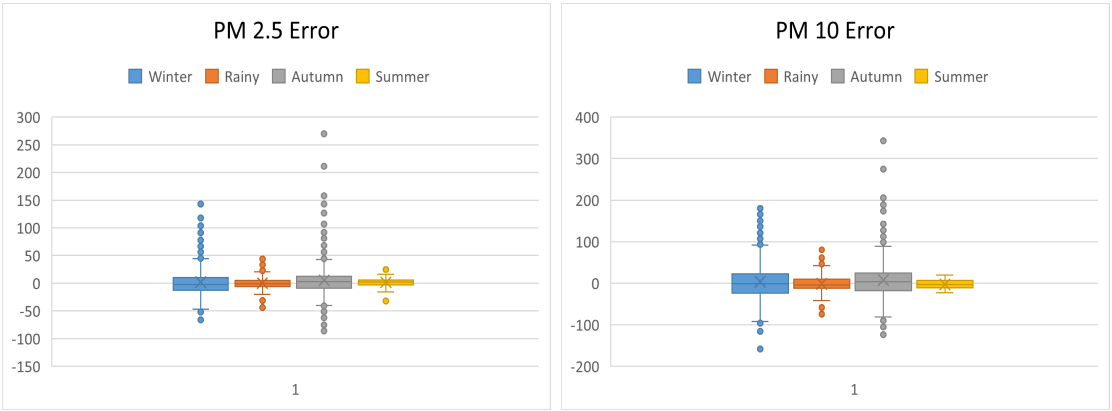


Figure 4.17: Seasonal error distribution for (a) $PM_{2.5}$ and (b) PM_{10} respectively.

Fig. 4.17 visualizes the distribution of errors between measured and estimated PM concentration values using a box plot. The medians are observed to be close to zero, reinforcing the fact that anomalous cases affect the mean(as observed in Table 4.8) but not the median, which is robust to outliers. The high average estimated biases and relative biases in the Autumn season, as seen in Table 4.8 appear to be due to the bursting of firecrackers during Diwali and lower temperatures during Winter. The high variability of the errors during these two seasons is also evident from Fig. 4.17.

4.5.14 Analysis of Results and Ablation Study

To study the effect of each of the sub-modules used in the proposed model, we conduct an ablation study to analyze the results generated by each module and their combinations with the actual concentration values. Similar system configuration and trained data have been used for conducting an ablation study for each module to avoid bias. This was necessary as the dataset is imbalanced due to the predominance of images with $PM_{2.5}$ and PM_{10} within the range of 150 $\mu g/m^3$ and 400 $\mu g/m^3$ respectively. Table 4.5 shows the performance of individual modules(High-Low Frequency and Depth Structural) as well as the combination of them, justifying the use of both modules together to get satisfactory results.

Table 4.7 shows the comparison of our model with two state-of-the-art CNN architectures. The modules used in the proposed model effectively extract the features responsible for the variation in the concentration of particulate matter and yield better results.

Four different scenarios have been employed to observe performance. Fig. 4.18 shows a scene at different times, Fig. 4.19 shows images with structural variations with the same or at different depths, Fig. 4.20 shows the same scene in different seasons and hence weather conditions and Fig. 4.21 shows images with varied light sources. The estimated concentration of $PM_{2.5}$ and PM_{10} as predicted by the proposed model Dark-AirNet is shown in Table 4.9 and 4.11 respectively. The predicted values are remarkably close to the actual ones for most of the cases. We have also shown results obtained by the baseline model Resnet-50 generating the next nearest results as shown in Table 4.10 and 4.12 respectively. The predicted values closest to the actual value have been displayed in bold font, ignoring minor differences.

Cases with a high difference between the actual and estimated values are mainly due to the skewness of the dataset, containing some images of regions having a very high value of $PM_{2.5}$ and PM_{10}. The visual information in such images seems to be less due to lack of proper light, heavy pollution, etc. Another confounding factor is the effect of rain on the visibility. Further, the absence of adequate light during low and high particulate matter concentrations has greatly enhanced the estimation error. It may be inferred that a judicious composition of the scene to be captured, having buildings at distances around 2.5 km, adequate lighting, but no blurring, would yield a good prediction by the model. The overall performance of the proposed model has already been presented in Table 4.5. However, a performance comparison could not be made as the only other work functions on a very limited dataset.

In Fig. 4.9 and 4.10 we have shown the plot of $PM_{2.5}$ and PM_{10} from July 2021

Fig. No.	***a***		***b***		***c***		***d***	
	Actual	Est.	Actual	Est.	Actual	Est.	Actual	Est.
4.18	105.50	**118.40**	128.60	**144.65**	143.40	**174.91**	192.00	**206.27**
4.19	44.00	**46.80**	44.00	57.46	44.00	**41.97**	44.00	**40.97**
4.20	58.20	**65.20**	13.10	**74.40**	127.50	**171.83**	277.70	186.96
4.21	82.40	**99.58**	53.30	**48.20**	129.10	**191.27**	62.60	**109.84**

Table 4.9: Estimated $PM_{2.5}$ predicted by the proposed model Dark-AirNet along with their actual value for images shown in the respective figure numbers in left to right format respectively.

Fig. No.	***a***		***b***		***c***		***d***	
	Actual	Est.	Actual	Est.	Actual	Est.	Actual	Est.
4.18	105.50	138.01	128.60	100.69	143.40	127.75	192.00	173.69
4.19	44.00	27.21	44.00	**52.4**	44.00	47.90	44.00	50.96
4.20	58.20	70.42	13.10	82.37	127.50	149.18	277.70	**237.33**
4.21	82.40	**99.90**	53.30	66.12	129.10	265.99	62.60	**109.06**

Table 4.10: Estimated $PM_{2.5}$ predicted by Resnet-50 model along with their actual value for images shown in the respective figure numbers in left to right format respectively.

to April 2022. The dotted line refers to the actual value, while the predicted one is shown with a dash-dot line. The major difference mainly occurred in the high-range concentration, as most of the time the concentration lies less than 200 (for $PM_{2.5}$) and 400 (for PM_{10}). The highest difference occurred during the Diwali celebrations on 4th November 2021 because of firecrackers, a rare event that occurred only on a day in the year. In Fig. 4.20(b), we have shown an image taken on 6th December 2021 where rainfall has caused the particulate concentration to drop but has also adversely affected the visibility. This confused the model, which estimated a higher value than the actual one. Similarly, in Fig. 4.21(c), we have displayed an image taken on 2nd Feb 2022 with poor lighting conditions leading to inaccurate estimation.

High particulate concentration typically remains within moderate levels, with exceptions only during the winter when the air becomes heavy and low wind speeds force the particulate matter to settle near the surface. Thus, a small percentage of images from the entire dataset corresponded to high or even very high PM concentrations. This tends to add some bias during the model training. A proper training dataset combined with model architecture and the Huber loss function helps minimize the differences.

Fig. No.	*a*		*b*		*c*		*d*	
	Actual	Est.	Actual	Est.	Actual	Est.	Actual	Est.
4.18	312.50	**322.73**	329.60	**274.24**	330.40	**323.11**	348.60	**343.47**
4.19	108.70	**112.67**	108.70	131.73	108.70	85.04	108.70	**97.18**
4.20	143.90	**154.39**	71.30	**145.19**	214.30	**266.38**	584.50	**392.81**
4.21	188.90	**266.38**	95.40	**119.56**	336.30	308.87	273.70	**295.39**

Table 4.11: Estimated PM_{10} predicted by the proposed model Dark-AirNet along with their actual value for images shown in the respective figure numbers in left to right format respectively.

Fig. No.	*a*		*b*		*c*		*d*	
	Actual	Est.	Actual	Est.	Actual	Est.	Actual	Est.
4.18	312.50	283.55	329.60	246.74	330.40	286.72	348.60	312.90
4.19	108.70	69.90	108.70	**112.40**	108.70	**109.34**	108.70	121.78
4.20	143.90	125.71	71.30	150.55	214.30	283.68	584.50	373.90
4.21	188.90	229.80	95.40	124.88	336.30	**331.06**	273.70	241.50

Table 4.12: Estimated PM_{10} predicted by Resnet-50 model along with their actual value for images shown in the respective figure numbers in left to right format respectively.

4.6 Concluding Remarks

The proposed model based on Convolutional Neural Networks (CNN) for predicting the concentration of $PM_{2.5}$ and PM_{10} is data-driven and independent of any spatial or temporal variability. Our model accounts for the scattering of light with changing particulate concentrations, which exhibit diurnal and seasonal variations. The model also considers the effects of structures and buildings at different depths from the measuring point, allowing for more accurate predictions in varying environments

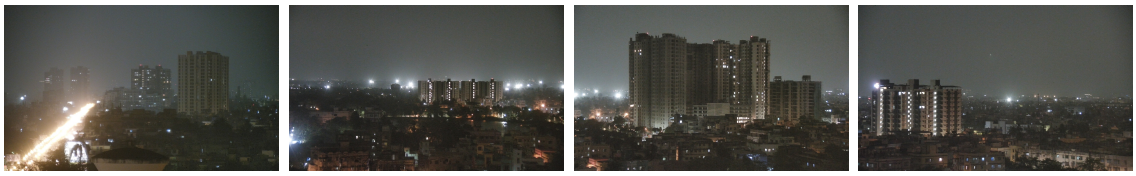
A diverse dataset of 60,000 nighttime images captured at regular intervals of 15 minutes, spanning from July 2021 to April 2022, with corresponding $PM_{2.5}$ and PM_{10} concentrations. The dataset captures a wide range of concentrations, from very low to very high levels, and covers scenes spanning several kilometers around the monitoring station. This rich dataset helps the model incorporate various meteorological factors that influence air quality, improving its robustness and predictive capabilities.

The proposed model, with its two-pronged approach of having a network to detect the frequency-based features and another one to construct the depth map, is well-suited for carrying out the enhancement of underwater images. Such images suffer from a loss of quality on account of various factors. Turbulences in the water arising



a b c d

Figure 4.18: Images taken on 02-Dec-2021, at 07:02, 09:04, 09:49, and 10:44 PM respectively.



a b c d

Figure 4.19: Images having structural variations taken on 01-Apr-2022, at 10:41, 10:46, 10:48, and 10:51 PM respectively.

from the passage of aquatic creatures or vehicles, the composition of the water, and the distance below the water surface, since attenuation of the various wavelengths constituting light is governed by it. Enhancement of underwater images is considered in the next chapter, using a transformer-based approach.



a b c d

Figure 4.20: Images taken during summer(1st Apr 2022), rainy(6th Dec 2021), autumn(3rd Nov 2021), and winter(1st Feb 2022) respectively.



a b c d

Figure 4.21: Images captured under different light conditions on 3rd Nov 2021, 6th Oct 2021, 2nd Feb 2022 and 10th Feb 2022 respectively.

Chapter 5

Aquaformer: Underwater Image Enhancement via Adaptive Transformer

5.1 Introduction

The restoration and enhancement of degraded underwater images have gained significant attention in recent years due to increased deep-sea exploration, marine biology research, and underwater diving activities. The underwater environment presents a complex ecosystem with many flora and fauna. However, the lighting conditions in these environments pose significant challenges due to the scattering and selective absorption of light. As the depth of water increases, the fading of colors becomes more pronounced, making it difficult to distinguish objects, especially at depths greater than 10 meters. This results in reduced visibility, color attenuation, and contrast loss, all imposing major limitations on underwater imaging applications. Therefore, restoring underwater images' original colors, clarity, and visibility is crucial for these applications.

Recent advancements in deep learning have introduced models based on convolutional neural networks (CNNs) and transformers. The availability of benchmark datasets containing pairs of degraded and clear underwater images has significantly improved the performance of data-driven models compared to traditional model-based approaches. Vision Transformers [21] have demonstrated superior performance over CNN-based models in many computer vision tasks. However, the effectiveness of Transformers in enhancing underwater images remains an under-explored area.

In this context, we propose a Transformer-based model with SwinTransformer[30] as the backbone to perform underwater image enhancement, to outperform CNN-based methods. Our proposed model aims to correct color distortions and produce clear, sharp images by introducing key modifications to various components of the Transformer architecture.

The chapter is organized into six sections. Section 5.2 presents a brief survey of related works. Section 5.3 provides a detailed discussion of the proposed model's approach. Finally, Section 5.4 describes the datasets used to train and test the model, along with the objective and subjective evaluation metrics. A detailed ablation study and a comparison with various state-of-the-art methods are also presented in this section.

5.2 Related Work

State-of-the-art methods for underwater image enhancement can be classified into three main categories: early physical models, models based on visual prior information, and data-driven models.

5.2.1 Physical Models

In this approach, a physical imaging model is established with a prior condition that is satisfied by key parameters, which form the foundation of the imaging model. Afterward, an inverse operation is performed on the degraded image to retrieve the original clear image.

The basis of a physical model-based enhancement technique is the use of prior information. For instance, Drews Jr et al. [125] applied the concept of Dark Channel Prior (DCP) to enhance images affected by underwater mediums. Other works utilizing prior knowledge include Color Attenuation Prior [126] and Minimum Information Loss Prior [127]. Akkaynak et al. [128] proposed a modified physical imaging model that leverages depth information and color channels to recover clear underwater images. However, estimating the depth map in an underwater environment remains a challenging task, often leading to color cast and artifacts in the corrected image. Zhang et al. [129] introduced a method based on minimal color loss and adaptive contrast enhancement performed locally. Zhuang et al. [130] developed a retinex variational model to enhance underwater images using hyper-Laplacian reflectance priors.

While these methods can yield reasonable results when input images align with the priors and parameters of the selected physical imaging model, they impose signif-

icant limitations on the model's scalability and robustness across different scenarios. Furthermore, they do not account for the subjective nature of human visual perception, resulting in poor visual quality in the restored images.

5.2.2 Visual Prior Models

The premise of these models is to modify pixel values to restore the original image. This involves techniques that use single or multiple image-based features to adjust the pixel values. Examples of such methods include Adaptive Contrast Enhancement [131] and White Balancing [132]. Some methods go a step further by fusing multiple features to create a clearer image. For instance, Ancuti et al. [133] proposed a weighted fusion of white-balanced and global contrast adjustments to enhance underwater images. Similarly, Li et al. [134] introduced a hybrid fusion approach that combines an improved white balance technique with contrast and saturation adjustments for image enhancement.

However, these methods fail to address the inconsistent attenuation in the various color channels, making them less effective on real underwater images, especially those captured in complex environments.

5.2.3 Data-driven Models

Advancements in deep learning have led to the development of several data-driven models that train on multiple datasets to directly estimate parameters and retrieve the original clear image based on these trainable parameters.

For instance, Li et al. [135] proposed WaterGAN, a model that generates underwater-like textured images from atmospheric images and depth maps in an unsupervised manner. The generated dataset is then used to train the network. A benchmark dataset containing large-scale realistic underwater images was introduced, along with a fusion network called Water-Net [9]. Additionally, Li et al. [136] proposed U-color, a transmission-based multi-color space embedding method that utilizes information from multiple color spaces and a reverse transmission map to enhance images.

Qi et al. [10] introduced SGUIE-Net, which applies both global and local enhancement strategies based on semantic information extraction through segmentation. However, the method heavily relies on accurate segmentation, a task that is difficult to perform with degraded images, leading to suboptimal results. Fu et al. [137] developed a probabilistic network that determines the enhancement distribution for degraded underwater images. Peng et al. [138] proposed U-transformer, a hybrid

model that combines a U-net CNN architecture with a multi-scale channel and spatial Transformer to enhance underwater images. Despite these models improving the visual quality of degraded images to some extent, they still fall short in more complex scenes.

In contrast, our proposed model focuses on retrieving high-quality color images by appropriately learning the features that mitigate the attenuation effects of each color channel while preserving the detailed information in the original image.

5.3 Proposed Model

5.3.1 An overview

The proposed model, AquaFormer, as shown in Fig. 5.1, is based on the SwinTransformer architecture as proposed in the work by Liu et al. [30], but with several key modifications that enhance its ability to retrieve high-quality underwater images from degraded ones. These modifications are designed to address specific challenges in underwater image restoration, such as color attenuation, loss of contrast, and the need to preserve fine details. Key features of the AquaFormer architecture include:

- Modification of the Norm Layer: The standard normalization layers have been altered to better adapt to the specific characteristics of underwater images, which often involve color distortions and low visibility.
- Non-linear Activation Function: The activation function has been adjusted to ensure that the model can effectively capture the complex relationships between different image features, especially under challenging underwater conditions.
- Aggregation Operation: Changes have been made to how features are aggregated to ensure that the model can combine information from different parts of the image effectively, improving its ability to restore the image quality.
- Adaptive Fusion Approach: Unlike traditional concatenation methods, an adaptive fusion approach has been employed to combine features from various scales. This allows the model to dynamically select and integrate the most relevant features at each scale, leading to better performance in terms of image clarity and detail.

For training, the model is optimized using the L2 loss function on a pair of raw degraded images and their corresponding clear reference images. This loss function

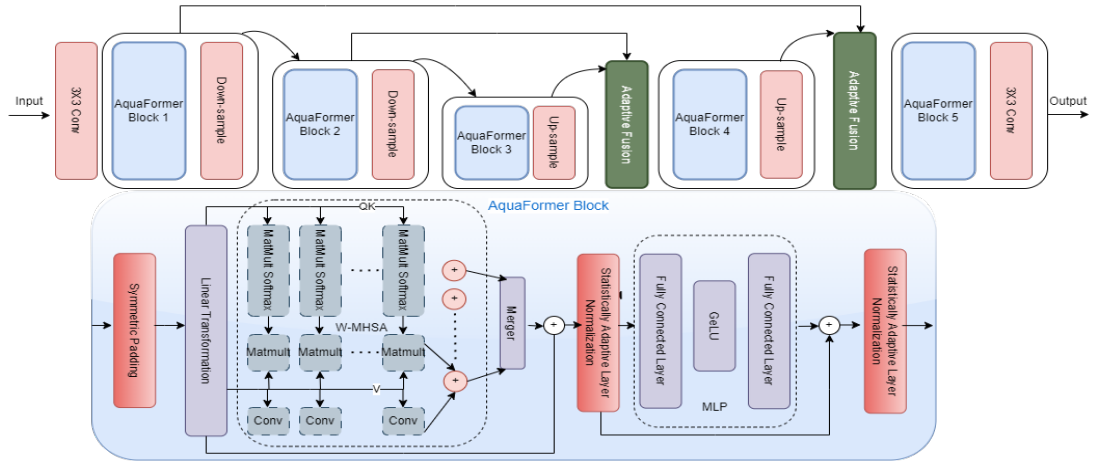


Figure 5.1: Architecture of a 5-stage Aquaformer with a detailed presentation of an Aquaformer Block

helps the model learn to minimize the pixel-wise difference between the predicted and ground-truth clear images, ensuring that the restored images are as close as possible to their true, clear counterparts.

The architecture and the modifications introduced in AquaFormer make it well-suited for the task of underwater image enhancement, enabling it to handle the unique challenges posed by underwater imaging environments.

5.3.2 Statistical Adaptive Layer Normalization

The normalization layer stabilizes the network training, enabling a smoother gradient, faster training, and better generalization accuracy. In a transformer, Layernorm [139] has been the most commonly used norm layer. Considering $X \in R^{b \times n \times c}$ where n = $h \times w$, (h = height and w = width) with c channels, the normalization process is expressed as:

$$Y = \gamma . \frac{X - \mu}{\sigma} + \beta \tag{5.3.1}$$

Here μ and σ refer to the statistical parameters mean and standard deviation of X, γ, and $\beta \in R^{b \times 1 \times c}$ are learnable scaling factors and biases respectively. In LayerNorm, μ and σ are computed along the c-axis, making $\{\mu, \sigma\} \in R^{b \times 1 \times c}$. The statistical parameters in an image help gauge its brightness and darkness. An image with a high mean and low standard deviation will appear very bright, while an image with a low mean and high standard deviation will appear very dark. Such normalization can often cause loss of color and contrast in the image. Thus, changes made in the Lay-

erNorm can reduce the negative effect that the normalization method adds to Y. We introduce learnable parameter $\{W_\mu, W_\sigma\} \in R^{b\times 1\times c}$ weighted with μ and σ respectively as in Eq. (5.3.2). We provide a small value ϵ to σ to maintain the stability. The modified layer normalization version helps our model preserve the statistical parameters μ and σ wherever necessary, depending upon the images, and thus helps the model in color restoration.

$$\hat{X} = \gamma \frac{(X - W_\mu \mu)}{\sqrt{(W_\sigma \sigma)^2 + \epsilon}} + \beta \tag{5.3.2}$$

Thus, our modified norm layer overcomes the problem of removing the mean and standard deviation from the input features.

5.3.3 Symmetric Padding on Shifted Window

One of the key innovations in SwinTransformer is the use of cyclic shifts combined with the Multi-Head Self-Attention (MHSA) mechanism. In SwinTransformer, cyclic shifts are applied to the image patches, and a masking approach is used to prevent attention from being paid to unrelated patches within a window. This technique helps to reduce computational complexity and improve the model's ability to capture local features.

In our proposed model, AquaFormer, we replace the cyclic shift-based masking approach with symmetric padding on the edge windows. This alteration prevents the interaction between unrelated patches within a window during the MHSA process, eliminating the need for a masking strategy as used in SwinTransformer. Symmetric padding ensures that the patches at the edges of the image are smoothly extended, thus maintaining the spatial integrity of the original image.

The use of symmetric padding provides two main advantages:

1. Avoidance of Masking: By performing padding instead of cyclic shifts, we prevent potential information loss that could arise from masking and ensure that each patch can interact effectively with its neighboring patches, even at the edges.

2. Preservation of Edge Details: Symmetric padding helps preserve fine details at the edges of the image by creating a smooth transition to the padded areas. This smooth transition ensures that the model does not introduce unnatural artifacts or distortions at the image boundaries, which is crucial for underwater image enhancement, where clarity and detail are vital.

Overall, this modification improves the model's ability to handle edge details and ensures that important information is preserved during the self-attention process.

5.3.4 Adaptive Fusion

In our proposed model, we perform an adaptive fusion of the features extracted by the encoder and the decoder to generate the final enhanced output. The fusion process is designed to effectively combine the low-level and high-level features extracted at different stages of the network. The process is as follows:

- Learnable Scalars: We introduce two learnable scalars, S_1 and S_2, which are added to the paths of the encoded and decoded features, respectively. These scalars help the model adaptively control the influence of the encoded and decoded features during the fusion process.

- Feature Fusion: After the addition of the learnable scalars, we perform a pixel-wise addition of the encoded and decoded features, which allows the model to merge the information from both paths effectively.

- Adaptive Average Pooling: The resultant feature map is then processed through an adaptive average pooling operation. This pooling operation helps to compress the feature map while retaining the essential spatial information, providing a more compact representation of the image that can be further processed.

- Feed-forward Network: The pooled output is passed through a 2-layer feed-forward network with a GELU activation function in between the layers. This network learns to refine the fused features further before generating the final result.

- Sigmoid Activation: After the feed-forward network, a sigmoid activation is applied to produce the final output, which gives the learnable scalar S_1 and its complement, $1 - S_1$, as S_2.

This adaptive fusion approach ensures that the model learns to combine the low-level and high-level features optimally, and the learnable scalars S_1 and S_2 allow the model to dynamically adjust the contribution of each feature path to the final output.

The fusion approach is visually depicted in Fig. 5.2.

5.3.5 Convolved W-MHSA mechanism

To preserve fine details while aggregating the input features, we modify the standard Softmax function that usually produces only positive weights, which tend to act like

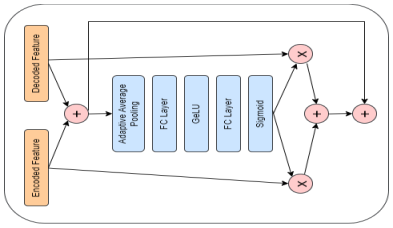


Figure 5.2: Adaptive fusion approach

a smoothing operation. Instead of directly applying Softmax, we introduce a convolution operation on the input features to capture local spatial information more effectively. Additionally, the features extracted by this convolution operation are fused with those obtained from the window-based multi-head self-attention (MHSA) mechanism.

This modified aggregation scheme helps to reduce the smoothing effect that Softmax typically imposes on patches and allows for the retention of important local details. The process can be mathematically expressed as in Eq. (5.3.3), where the convolution operation and the MHSA are integrated to compute the attention more effectively.

By performing this aggregation in the modified way, the model can better capture both local spatial dependencies and the broader contextual information provided by the attention mechanism. This combination helps to improve the quality of the image restoration while maintaining important details.

The modified aggregation scheme can be summarized as:

$$Attn(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V + Conv(V) \tag{5.3.3}$$

Here, d is its feature dimension and Conv is the sequential block consisting of the operations [conv-ReLU-conv].

5.4 Implementation Details with Experimental Data

This section details the experimental setup used to develop and evaluate our model. It encompasses Datasets: The datasets employed for training and testing the model are discussed, highlighting the diversity of images and conditions considered for the experiments; Evaluation Metrics: We introduce the performance metrics used to assess the quality of the enhanced images. These metrics are crucial for both subjective

and objective comparisons with existing methods; Model Implementation: The technical specifications and implementation details of our model, including the hardware and software environment, training procedures, and hyperparameters, are provided to ensure reproducibility of the results; Ablation Study: An ablation study is conducted to investigate the contribution of different components of our model. This helps in understanding the significance of each design choice and how they collectively contribute to the model's performance; Comparison with State-of-the-Art: Our model's performance is compared with several state-of-the-art methods to showcase its competitive edge in terms of both qualitative and quantitative results. We evaluate the robustness of the model using various assessments, including visual quality and numerical metrics.

5.4.1 Implementation Details

The AquaFormer model is designed with a series of five transformer blocks, each with specific configurations tailored to enhance the model's ability to process and recover high-quality underwater images. The details of the block settings are as follows:

1. Embedding Dimensions: The embedding dimensions for each block are set as {48, 96, 192, 96, 48}

2. Multi-Head Attention: The number of multi-heads for each block is {2, 4, 6, 4, 2}.

3. Block Depth: The depth for each block is consistently set to {16, 16, 16, 16, 16}, providing a uniform structure across all blocks.

4. Patch Size: Each input patch is set to a size of 128×128 pixels.

5. Batch Size: The batch size for training is set to 32.

For optimization, the model utilizes the Adam optimizer coupled with a cosine annealing strategy. The training process runs for 500 epochs, with the initial learning rate set to 1×10^{-6}. This configuration ensures the model trains effectively and converges on optimal performance.

The implementation is done using the PyTorch framework, and the model is simulated on an Nvidia RTX A5000 GPU, ensuring efficient training and inference processing.

5.4.2 Datasets

We conduct extensive testing and validation of our proposed AquaFormer model using a variety of benchmark datasets to evaluate its robustness and performance. Our model is trained using 800 raw-reference pair-images from the UIEB dataset [9] and 1000 synthetic pair-images from [12], where the synthetic dataset contains 10 categories corresponding to 10 different types of water, divided into two groups: I, IA, IB, II, and III for open ocean water, and 1, 3, 5, 7, and 9 for coastal water.

To test the robustness of the model, we adopted the testing strategy from [136], conducting experiments across several benchmark datasets. These datasets include:

1. Test-R90: This dataset includes 90 pairs of underwater images randomly selected from the UIEB dataset [9].

2. Test-S1000: Comprising 1000 synthetic underwater image pairs (100 pairs from each of the 10 water categories from [12]).

3. Test-C60: Includes 60 real underwater images from the UIEB dataset with no reference images, which are considered particularly challenging for state-of-the-art methods.

4. Color-Checker7: Contains seven underwater images featuring a color-checker, taken with different camera settings as described in [14], to evaluate the robustness of color correction methods.

5. RUIE: A dataset used to evaluate methods on real-world underwater images, sourced from [140].

6. EUVP: Includes 11,435 paired underwater images, as well as nearly 6000 unpaired images in three subsets. The evaluation for EUVP was performed using the first 10 images from each subfolder, following the methodology outlined in [10].

7. SQUID: Contains 57 underwater stereo image pairs captured at four different dive sites in Israel, each with a resolution of 1827×2737. Due to the large size of the images, we used the low-resolution version of the dataset provided in [136].

8. SUIM-E: This dataset includes 1635 real underwater images with pixel-level annotations for eight object categories.

By leveraging these diverse datasets, we evaluate the model on underwater imaging scenarios, ensuring that the AquaFormer model performs robustly across different conditions and image qualities. The model's effectiveness is validated through both quantitative and qualitative assessments, demonstrating its ability to restore high-quality, color-corrected underwater images.

5.4.3 Evaluation Metrics

We assess the performance of the proposed AquaFormer model using a combination of full-reference and no-reference evaluation metrics. For full-reference metrics, we compute the following:

1. Peak Signal-to-Noise Ratio (PSNR): Measures the quality of the enhanced image by comparing it with the ground truth image, where a higher PSNR indicates better image quality.

2. Structural Similarity Index (SSIM): Evaluates the similarity between the enhanced and reference images by considering luminance, contrast, and structure. Higher SSIM values suggest better quality and perceptual similarity to the reference.

For no-reference metrics, which do not require a ground truth image, we use the following:

1. Blind Image Spatial Quality Evaluator (BRISQUE): An image quality assessment metric that evaluates the naturalness of the image based on spatial domain features.

2. Naturalness Image Quality Evaluator (NIQE): Assesses the perceptual quality of an image based on natural scene statistics, with lower values indicating better image quality.

3. Perception-based Image Quality Evaluator (PIQE): A metric that predicts image quality based on perceptual quality and human visual system properties, where lower values indicate better quality.

These evaluation metrics allow us to measure the effectiveness of our model in both scenarios, whether or not reference images are available for comparison, providing a comprehensive understanding of its performance.

5.4.4 Comparison of performance

We have benchmarked the performance of the proposed AquaFormer model against seven of the most recent state-of-the-art methods in underwater image enhancement. These methods, which span both traditional and neural network-based image enhancement techniques, are: WaterNet[9], Ucolor[136], MLLE [129], SGUIE-Net [10], PUIE-Net[137], HLRP [130], and UTransformer[138].

5.4.5 Color Restoration Performance

We have used the SQUID and ColorChecker7 datasets to analyze the color restoration performance. Pre-trained models for all the state-of-the-art methods that are based on deep learning techniques have been used for comparison.

The SQUID dataset provides diverse underwater images with varying depths (from 3 meters to 30 meters). In Fig. 5.4, we present the enhancement results obtained by various state-of-the-art methods and the proposed Aquaformer on the SQUID dataset. MLLE and HLRP do remove the underwater effects from the images, but add color-cast and artifacts in their respective outputs. SGUIE-Net was unable to remove the effect of being underwater from the images, while PUIE-Net improved the output to some extent by removing the water effect. The results obtained by our proposed model not only prevent the color-cast problem but also remove the underwater effect from the images. To evaluate the accuracy of color correction, we compute the average value of NIQE, BRISQUE, and PIQE metrics as shown in Table 5.3. The results show competitive output compared to the state-of-the-art methods. It may be noted that while MLLE exhibits the best performance for the Test-C60 dataset with respect to metrics BRISQUE and PIQE, it adds color-cast and artifacts in the resultant images. The metrics are unable to capture such effects, showing deficiencies in the metrics themselves.

The ColorChecker7 dataset presents seven underwater images, each with a color checkerboard taken from seven cameras with different specifications. The dataset is used to check the robustness of various state-of-the-art methods in terms of verifying color corrections in underwater images. In Fig. 5.4, we have covered images taken from two different cameras (Fuji Z33 and Olympus T6000). MLLE and HLRP make the enhanced outputs over-contrasted. SGUIE-Net, PUIE-Net, and UFormer make some improvements in color correction. We quantify the performance by computing the CIEDE2000 scores for all the color checkerboard images in the dataset obtained by state-of-the-art and our method, as displayed in Table 5.3.

5.4.6 Visual Quality Performance

We have used UIEB and UIEB Challenging, Synthetic, EUVP, and SUIM-E datasets to make qualitative and quantitative assessments of visual quality performance. HLRP and MLLE methods in the majority of the cases introduce color deviations, adding artifacts, and giving the highest weightage to the red channel. SGUIE-Net makes good color correction but adds constraints of performing region-wise enhancement through segmentation, which may not be accurate due to the loss of semantic information. Further, it resizes the input image to a resolution of 640×480, leading to a loss of detailed information. In many cases, such as severely degraded images from [12], the extraction of semantic information is extremely challenging due to inaccurate segmentation. Thus, existing approaches fail in many such cases, adding artifacts and color casts to the output. Notably, the Aquaformer not only performs robust color correction but also prevents the addition of color cast and artifacts in the resultant images.

Fig. 5.3 shows the results obtained from the UIEB Test-R90, SUIM-E, EUVP, and Test S-1000 test datasets and compares them with the results obtained from state-of-the-art methods. In Table 5.2, we quantify the results by computing the average value of the evaluation metrics obtained for the multiple Test datasets.

It may be observed that some of the methods used for comparison failed to make color corrections, while some added color-cast in the resultant images. The proposed Aquaformer performs the enhancement efficiently even in cases that show severe degradation, performing proper color correction and preserving the detailed information of the input image.

In Fig. 5.4, results obtained from ColorChecker7, SQUID, and UIEB Test-C60 [9] challenging datasets have been presented. The proposed model consistently shows robust performance vis-à-vis other state-of-the-art methods. Table 5.3 qualitatively evaluates the experimental results based on the average values of NIQE, BRISQUE, and PIQE metrics.SQUID dataset contain images that are of very high resolution. The previous works have resized the image to let their model generate the output which not only tried to enhance the image but also amplified that noise that got added due to resizing. In our model, we have cropped the image into patches each of size 256x256 and concatenated them along the batch dimension. Each patch is enhanced and the final output is concatenated to its original dimension. leading to no loss of data or noise amplification. This led to the superior recording lowest values in all no-reference metrics compared to the SOTA. Assessments based on these metrics show that traditional models such as MLLE give higher scores. However, the visual perception of the resultant images is inferior as seen in Fig. 5.3 and 5.4 covering various benchmark datasets.

Since these metrics use contrast as one of their parameters, over-contrast of the resultant output enhances the score. Our model in such cases gives superior performance compared to traditional image enhancement techniques. This validates that the model effectively works for underwater images, making proper color corrections and visual enhancements.

5.4.7 Ablation Study

We perform an ablation study of our proposed model to substantiate the effectiveness of many modifications that made our model highly robust to diverse underwater scenes. We have conducted experiments under the following conditions:

- Impact of statistical adaptive layer normalization against layer normalization.
- Symmetric padding vs cyclic shift
- Adaptive fusion vs normal fusion.
- Modified MHSA vs normal MHSA

We have been able to present only a small portion of the ablation results due to paucity of space, in Table 5.1. The PSNR(P) and SSIM(S) metrics obtained using the models already stated on the raw images from two benchmark datasets, Test-R90 and SUIM-E test, are displayed. This demonstrates the effectiveness of the modifications incorporated into the proposed model.

Table 5.1: Ablation Study for judging the effectiveness of the contributions, conducted on Test-R90 and SUIM-E

Study Area	Test-R90		SUIM-E Test	
	P (dB)	S	P (dB)	S
Layer Normalization	21.82	0.790	24.23	0.855
Cyclic Shift	20.93	0.772	22.31	0.817
Normal fusion	20.87	0.770	22.56	0.806
MHSA	21.56	0.778	23.15	0.826
Aquaformer (all 4)	23.61	0.856	26.81	0.939

5.5 Concluding Remarks

The adaptation of the window-based modified self-attention mechanism that combines the impact of convolution on the features with that of the attention mechanism



Figure 5.3: Visual quality performance comparison on images (row-wise) with the first three from Test-R90[9] followed by the next three from SUIM-E[10], two from EUVP[11], and the last two from Test-S1000[12] test dataset. Zoom to the maximum extent possible to get full clarity.

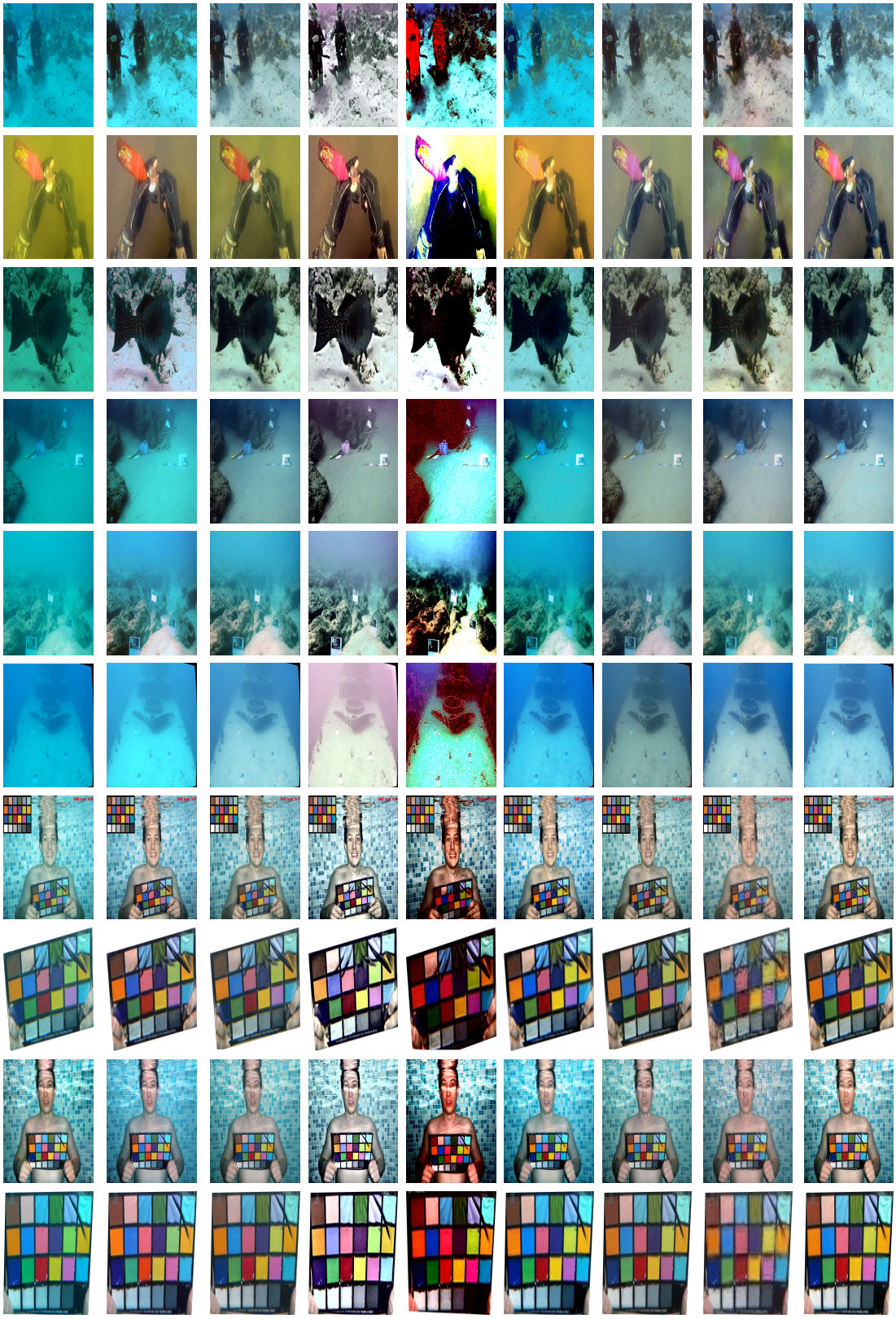


Figure 5.4: Visual quality performance comparison (row-wise images) with the first three images from Test-C60[9], the next three from SQUID[13], and the last two image-colorchecker pair from Color-Checker7[14] dataset. Zoom to the maximum extent possible to get full clarity. The smaller the CIEDE score, the better the result.

Method	Test-R90		SUIM-E Test		EUVP Test		Test-S1000	
	P (dB)↑	S↑	P (dB)↑	S↑	P (dB)↑	S↑	P (dB)↑	S↑
WaterNet	21.03	0.764	21.16	0.824	19.32	0.653	14.40	0.549
Ucolor	21.25	0.770	20.13	0.734	20.16	0.748	23.25	0.823
MLLE	17.99	0.637	17.33	0.658	15.11	0.572	13.16	0.481
HLRP	12.62	0.360	12.35	0.355	11.32	0.237	9.59	0.205
SGUIE	20.02	0.700	21.09	0.826	18.98	0.651	14.14	0.381
PUIE	21.37	0.760	20.88	0.830	19.02	0.686	14.52	0.343
UFormer	20.15	0.719	19.89	0.726	25.36	0.874	14.40	0.350
Aquaformer	23.61	0.856	26.81	0.939	20.04	0.744	24.94	0.850

Table 5.2: Quantitative comparison among SOTA methods using reference testing set. The top three scores of PSNR(P) and SSIM(S) in descending order are marked with red, green, and blue respectively

Method	Test-C60			SQUID		
	B↓	N↓	P↓	B↓	N↓	P↓
WaterNet	33.67	4.14	48.01	51.20	5.90	91.83
Ucolor	52.64	5.74	85.62	51.89	5.98	90.28
MLLE	27.31	4.16	39.74	54.05	5.62	88.12
HLRP	37.23	4.27	58.48	53.78	5.23	87.90
SGUIE	43.53	4.77	72.18	51.58	5.87	89.95
PUIE	35.89	4.09	51.01	51.48	5.92	88.45
UFormer	52.95	5.69	85.61	48.38	6.44	98.90
Aquaformer	33.13	4.06	42.91	9.72	3.08	18.05

Table 5.3: Qualitative comparison among SOTA methods. The top three scores of BRISQUE(B), NIQE(N), and PIQE(P) in ascending order are marked with red, green, and blue respectively

is a vital innovation in the work. The window-based approach helped model to perform enhancement of images of any size, thus overcoming the problem of resizing of large size images. The effect is clearly visible on the SQUID dataset where the model preserved every details of the image while restoring the colors which the SOTA methods failed.

The incorporation of the Statistical Adaptive Layer Normalization is responsible for preserving the statistical parameters of each channel of the patch. This helps maintain the underwater naturalness of the image while making color corrections.

The introduction of symmetric padding to maintain the cross-window connection that the Transformer uses to incorporate attention between windows has been very effective in reducing border effects. This has helped in avoiding any artifacts and

distortions that may be added due to masked cyclic shifts and preserving the original content.

A notable feature of the proposed approach is the adaptive feature fusion approach instead of simple concatenation, which utilizes features generated from different scales of the input features.

The combination of all such adaptations and innovation helped the model produce superior results compared to SOTA methods.

Chapter 6

Conclusions

6.1 A Summary of the work done

Approaches for restoring and enhancing images degraded by environmental factors and applications based on these have been developed. Environmental factors considered range from fog or haze in the atmosphere for images acquired out in the air to local disturbances or murkiness in the case of images taken underwater. The underlying principles and methodology form an entire gamut, ranging from employing sophisticated hand-crafted features, adoption of a hybrid between such features and machine learning and finally deep learning networks. A robust dataset consisting of both night-time and day-time images has been created for the task of accurately estimating the concentration of particulate matter, specifically $PM_{2.5}$ and PM_{10} from images. The images are analyzed to determine the impact of various meteorological factors responsible for the concentration of particulate matter in a region. To accurately estimate the concentrations, we have developed a model to predict the type of weather, namely sunny, cloudy, rainy, and hazy, from an image. We have also proposed a monochromatic depth map-generating model from a given color image, and both models are helpful in the accurate estimation of the particulate concentration.

Handcrafted features have been designed and employed in a pyramid fusion method for image dehazing. The algorithm uses two derived input images—contrast-enhanced and white-balanced—from which five feature maps are generated, namely chromaticity, colorfulness, entropy, saliency, and visibility. These feature maps help the method differentiate between haze-free and hazy images. The extracted features are processed in parallel: one stream passes through a Laplacian filter to detect edge boundaries, while the other stream is filtered using a Gaussian filter to remove noise. The features

are then fused at multiple scales by applying down-sampling operations, followed by a fusion step in the form of a feature pyramid. This fusion allows the extraction of valuable multi-scale features through up-sampling operations. A significant advantage of the proposed method is that it does not require GPU-based support for training, nor does it need the estimation of dehazing model parameters, such as the transmittance map and airlight. This makes the method lightweight and suitable for real-time applications. To evaluate its performance, we compare the proposed method with several state-of-the-art techniques using different benchmark datasets, highlighting its effectiveness and efficiency.

The handcrafted features designed in above chapter are next used in conjunction with machine learning to tackle two important real-life applications – determination of the type of weather from a give image and next, computing the particulate matter concentration present in the air from an image. For the former application, a fusion model has been proposed that integrates handcrafted and neural network-extracted features, which contribute to highly accurate weather prediction. To enhance the performance of our model, we introduce an Unet-based SKYNET model specifically for detecting and segmenting sky regions within the image. Additionally, we employ a multi-class latent joint support vector machine framework to make classification decisions based on a 605-dimensional feature vector extracted from the image. This approach allows the model to classify five distinct weather conditions: "Sunny", "Cloudy", "Haze", "Snow", and "Rainy".

In the second application, we have proposed a daytime image-based $PM_{2.5}$ concentration estimation method. Two different algorithms have been proposed, one adopts a RoI-based strategy employing two different Regions of Interest (RoIs) spaced at two locations widely apart about the camera while the other uses information from the entire image. For both algorithms, a powerful dehazing technique was deployed to generate a reference image for using the SSIM metric as a quantifier for assessing the level of $PM_{2.5}$. The results obtained using the ETR model are very promising and represent a significant improvement over the state-of-the-art in making a very good quality prediction of $PM_{2.5}$. Since the ROI-based model exhibits the best performance, it could be used for a situation where there exists the provision of RoI selection by the user. Hence this work has two major contributions: one is the selection of a well-designed feature set that is both unique and independent of any extraneous information source such as the Internet and second, the deployment of this set into a powerful model that can produce very good quality prediction of $PM_{2.5}$ from day-time images.

In order to come up with an effective solution for estimating the $PM_{2.5}$ concentra-

tion from images acquired at night, obtaining an accurate depth map of the region covered in the image is of paramount importance. A deep learning model for estimating the monocular depth of a given color image has been designed next. The novelty of the model lies in its ability to extract features using a deep learning network, which is further enhanced by Conditional Random Fields (CRF). Our novel supervised approach for estimating depth maps integrates regional structural constraints into an encoder-decoder architecture. By leveraging multi-scale information from various resolution layers and incorporating local feature adaptations through regional constraints, the model significantly improves depth accuracy. This method effectively addresses the limitations of existing approaches, which often neglect local properties that are critical for accurate depth estimation. Experimental results validate the effectiveness of the proposed model, showcasing its ability to outperform state-of-the-art methods across multiple benchmark datasets. These results highlight the substantial improvements in depth map accuracy and demonstrate the model's capacity to handle complex depth estimation tasks with high precision.

We next introduce a high-performing night-time image-based model designed to estimate the concentration of particulate matter in the air of a region. The model utilizes two distinct modules to achieve this task. The high-low frequency module captures the spectral reflectance of light affected by particulate matter. As particulate concentration increases, light scattering intensifies, reducing the amount of light reaching the surface and causing haze or dimming in the atmosphere. In contrast, lower particulate concentrations result in less scattering, allowing more light to reach the surface, leading to clearer skies and improved visibility. The depth module as elaborated in the chapter generates a depth profile from a color image, providing valuable insights into the proximity of various structures relative to the camera. Distant buildings show lower visibility, while closer structures are more easily discernible. Similarly, lower particulate concentrations result in clearer visibility, while higher concentrations lead to reduced visibility. Without depth information, the model may confuse low visibility due to distance with that caused by high particulate concentrations. This highlights the importance of depth information for making accurate particulate matter concentration estimates and reinforces its critical role in the model. A comprehensive dataset has been created, encompassing images captured in diverse weather conditions and locations, which helps the model handle a wide range of real-world scenarios.

Finally, a novel underwater image enhancement model is built on the standard SwinTransformer architecture with several crucial modifications. The modifications enable the model to generate far superior image enhancement outputs, where many

existing state-of-the-art methods fall short. The first major contribution is the incorporation of a statistically adaptive layer normalization, designed to preserve each channel's statistical parameters—an essential aspect for underwater image enhancement. This helps maintain the natural appearance of underwater scenes while correcting color distortions. The second key contribution is symmetric padding around the input boundaries. This approach ensures that the model retains important edge information during the enhancement process, preventing the loss of crucial spatial details that might otherwise be missed by other models. Furthermore, we introduce an adaptive fusion approach that dynamically combines the features extracted by the encoder and decoder, replacing the traditional concatenation method. This method allows for more effective integration of multi-scale features, leading to better image quality. The fourth contribution involves the window-based self-attention mechanism, a core component of SwinTransformer, which has been adapted to better handle the characteristics of underwater images. This adjustment improves the model's ability to focus on relevant patches and enhances the restoration of visual details.

These advancements collectively make the proposed model more robust and capable of handling a diverse range of underwater scenes, where many previous approaches do not perform well. Through extensive experiments on benchmark datasets, we demonstrate the superiority of our model in addressing color channel attenuation and preserving detailed image information.

6.2 Scope for Future Work

This thesis establishes a strong foundation for vision-based environmental perception across adverse imaging conditions; however, several promising research directions remain open for long-term investigation.

- The long-term goal of the image dehazing research is to develop a fully generalized, real-time, and depth-aware dehazing framework capable of handling dense, heterogeneous, and non-uniform atmospheric degradations across diverse outdoor environments. This includes building large-scale multi-weather dehazing datasets, integrating depth and 3D geometric priors, and designing transformer-based multimodal dehazing architectures. The ultimate objective is to deploy these models in autonomous driving systems, surveillance networks, and satellite imaging pipelines for reliable vision under all visibility conditions.

- The long-term objective of multi-weather detection is to realize a comprehen-

sive, multi-label weather understanding system that can recognize simultaneous weather phenomena in complex real-world scenes. Future developments aim at fusing visual data with real-time meteorological sensor data to achieve high-precision environmental awareness. A further long-term vision includes embedding this capability into smart-city infrastructure, traffic control systems, and early warning platforms for climate-resilient urban planning.

- The long-term goal of air particulate concentration estimation is to establish a unified, end-to-end, vision-based air quality monitoring framework that operates seamlessly in both daytime and nighttime conditions. This includes eliminating manual RoI selection, integrating multimodal meteorological variables, and enabling real-time predictive modeling through mobile and edge-device deployment. The ultimate goal is to contribute toward cost-effective, large-scale air quality monitoring systems for public health, environmental governance, and climate research.

- The long-term vision for night-time depth estimation is to achieve high-fidelity 3D scene reconstruction under extremely low-light and adverse weather conditions by integrating geometric priors, multimodal sensing, and neural radiance field (NeRF) representations. This research aims to support applications in autonomous navigation, disaster management, remote surveillance, and low-light robotics by enabling accurate spatial understanding under visually degraded environments.

- The model for monocular depth estimation produces high-quality depth maps, yet struggles in complex scenes, particularly those affected by severe weather conditions such as haze, sandstorms, and similar challenges. The incorporation of additional geometric features like shape, orientation, and spatial relationships, need to be explored. A promising avenue for future development is using neural radiance fields (NeRF) to generate 3D depth maps, enabling the creation of detailed 3D representations of objects or entire scenes, even in challenging environments.

- Extending the approaches developed, construction of a single model to determine the concentration of particulate matter for both daytime and nighttime images would be very useful. This could lead to developing a mobile API that can help in estimating the concentration in different areas. Additionally, the integration of meteorological factors will further improve the model's accuracy

and applicability. The incorporation of principles of Physics and Optics is also expected to improve performance.

- The long-term goal of underwater image restoration is to develop a unified, real-time, and multimodal underwater vision framework capable of supporting long-term marine exploration, habitat monitoring, and autonomous underwater navigation. This includes designing lightweight transformer architectures, integrating depth and optical sensor data, and enabling deployment on Autonomous Underwater Vehicles (AUV) and Remotely Operated Vehicles (ROV). The broader vision is to establish a robust visual intelligence system for sustainable ocean observation, underwater infrastructure inspection, and marine biodiversity conservation.

Bibliography

[1] J. El Khoury, J. B. Thomas, and A. Mansouri, "A database with reference for image dehazing evaluation," *Journal of Imaging Science and Technology*, vol. 62, 01 2017. xv, 27

[2] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-haze: A dehazing benchmark with real hazy and haze-free outdoor images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 867–8678. xv, 29, 30

[3] C. O. Ancuti, C. Ancuti, R. Timofte, and C. D. Vleeschouwer, "I-haze: a dehazing benchmark with real hazy and haze-free indoor images," in *arXiv:1804.05091v1*, 2018. xv, 29, 30

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. xvii, 84, 87

[5] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *CVPR*, 2018. xvii, 87, 88, 89

[6] C. Godard, O. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," 11 2019. xvii, 87, 88, 89

[7] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *International Conference on 3D Vision*, 2022. xvii, 74, 87, 88, 89

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban

scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223. xvii, 84, 88

[9] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2020. xviii, 105, 112, 114, 115, 117, 118

[10] Q. Qi, K. Li, H. Zheng, X. Gao, G. Hou, and K. Sun, "Sguie-net: Semantic attention guided underwater image enhancement with multi-scale perception," *IEEE Transactions on Image Processing*, vol. 31, pp. 6816–6830, 2022. xviii, 105, 112, 114, 117

[11] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227–3234, 2020. xviii, 117

[12] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognition*, vol. 98, p. 107038, 2020. xviii, 112, 115, 117

[13] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2822–2837, 2021. xviii, 118

[14] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 379–393, 2018. xviii, 112, 118

[15] A. Galdran, "Image dehazing by artificial multiple-exposure image fusion," *Signal Processing*, vol. 149, pp. 135–147, 2018. xix, 17, 30, 31

[16] M. Navale, P. Sawant, and P. Bagal, "Color attenuation prior (cap) for single image dehazing," *International Journal Of Engineering And Computer Science*, vol. 7, pp. 23 578–23 584, 02 2018. xix, 30, 31

[17] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 617–624. xix, 30, 31, 32

[18] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1674–1682. xix, 17, 30, 31

[19] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3888–3901, 2015. xix, 18, 30, 31

[20] W. E. K. Middleton, *Vision through the Atmosphere*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1957, pp. 254–287. 3

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010. 4, 103

[22] C. Mobley, *Light and Water: Radiative Transfer in Natural Waters*, 01 1994. 5, 55

[23] "Chapter 10 irradiance," in *Optical Oceanography*, ser. Elsevier Oceanography Series, N. Jerlov, Ed. Elsevier, 1968, vol. 5, pp. 115–132. 6

[24] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6723–6732. 6

[25] H. Bhandari and S. Palit, "Looking beyond the haze: A pyramid fusion approach," in *Neural Information Processing*. Cham: Springer International Publishing, 2021, pp. 654–665. 12, 58

[26] H. Bhandari, S. Chowdhury, and S. Palit, "A hybrid human-machine system for image-based multi-weather detection," in *Image and Vision Computing (IVCNZ)*. Cham: Springer Nature Switzerland, 2023, pp. 315–329. 12

[27] S. Chakraborty, S. Palit, and H. Bhandari, "Can an image tell the tale: Looking beyond the haze to determine pm2.5 concentration," in *2024 IEEE International Joint Conference on Neural Network (IJCNN), Yokohama, Japan*, 2024, pp. 1–8. 12

[28] H. Bhandari and S. Palit, "Real-net: A monochromatic depth estimation using regional attention and local feature mapping," in *Image and Video Technology*. Singapore: Springer Nature Singapore, 2024, pp. 302–311. 13

[29] H. Bhandari and S. Palit, "Aquaformer: Underwater image enhancement via adaptive transformer," in *2024 IEEE International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1993–1997. 13

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 9992–10 002. 13, 104, 106

[31] S. Fang, J. Zhan, Y. Cao, and R. Rao, "Improved single image dehazing using segmentation," in *2010 IEEE International Conference on Image Processing*, 2010, pp. 3589–3592. 15

[32] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013. 15

[33] H. Park, D. Park, D. K. Han, and H. Ko, "Single image haze removal using novel estimation of atmospheric light and transmission," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 4502–4506. 15

[34] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2011. 15, 17

[35] A. Cantor, "Optics of the atmosphere–scattering by molecules and particles," *IEEE Journal of Quantum Electronics*, vol. 14, no. 9, pp. 698–699, 1978. 16

[36] Y. Song, H. Luo, J. Ma, B. Hui, and Z. Chang, "Sky detection in hazy image," *Sensors*, vol. 18, no. 4, 2018. 16, 22

[37] S. Fang, X. Xia, X. Huo, and C. Chen, "Image dehazing using polarization effects of objects and airlight," *Opt. Express*, vol. 22, no. 16, pp. 19 523–19 537, Aug 2014. 17

[38] R. T. Tan, "Visibility in bad weather from a single image," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8. 17

[39] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, p. 1–9, aug 2008. 17

[40] D. Berman, T. Treibitz, and S. Avidan, "Single image dehazing using haze-lines," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 720–734, 2020. 17

[41] H. Israël and F. Kasten, *KOSCHMIEDERs Theorie der horizontalen Sichtweite*. Wiesbaden: VS Verlag für Sozialwissenschaften, 1959, pp. 7–10. 18

[42] J. El Khoury, J. B. Thomas, and A. Mansouri, "A database with reference for image dehazing evaluation," *Journal of Imaging Science and Technology*, vol. 62, 01 2017. 18

[43] C. O. Ancuti, C. Ancuti, and P. Bekaert, "Effective single image dehazing by fusion," in *2010 IEEE International Conference on Image Processing*, 2010, pp. 3541–3544. 18

[44] Y. Li, Q. Miao, R. Liu, J. Song, Y. Quan, and Y. Huang, "A multi-scale fusion scheme based on haze-relevant features for single image dehazing," *Neurocomput.*, vol. 283, no. C, p. 73–86, Mar. 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2017.12.046 18

[45] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3253–3261. 18

[46] M. Tan, T. Fang, Y. Fan, W. Wu, Q. She, and H. Gan, "Image-dehazing method based on the fusion coding of contours and colors," *IEEE Access*, vol. 7, pp. 147 857–147 871, 2019. 19

[47] X. Yuan, M. Ju, Z. Gu, and S. Wang, "An effective and robust single image dehazing method using the dark channel prior," *Information*, vol. 8, no. 2, p. 57, 2017. [Online]. Available: https://www.mdpi.com/2078-2489/8/2/57 22

[48] L. Zhang, Z. Gu, and H. Li, "Sdsp: A novel saliency detection method by combining simple priors," in *2013 IEEE International Conference on Image Processing*, 2013, pp. 171–175. 24

[49] N. Hautière, J.-P. Tarel, D. Aubert, and É. Dumont, "Blind contrast enhancement assessment by gradient ratioing at visible edges," *Image Analysis & Stereology*, vol. 27, no. 2, pp. 87–95, 2008. 29

[50] M. Roser and F. Moosmann, "Classification of weather situations on single color images," in *2008 IEEE Intelligent Vehicles Symposium*, 2008, pp. 798–803. 35, 36

[51] X. Yan, Y. Luo, and X. Zheng, "Weather recognition based on images captured by vision system in vehicle," in *Advances in Neural Networks – ISNN 2009*, W. Yu, H. He, and N. Zhang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 390–398. 35, 36

[52] K. G. Derpanis, M. Lecce, K. Daniilidis, and R. P. Wildes, "Dynamic scene understanding: The role of orientation features in space and time in scene classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1306–1313. 36

[53] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3726–3733. 36

[54] R. Baltenberger, M. Zhai, C. Greenwell, S. Workman, and N. Jacobs, "A fast method for estimating transient scene attributes," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–8. 36

[55] D. Glasner, P. Fua, T. Zickler, and L. Zelnik-Manor, "Hot or not: Exploring correlations between appearance and temperature," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3997–4005. 36

[56] M. Islam, N. Jacobs, H. Wu, and R. Souvenir, "Images+weather: Collection, validation, and refinement," 01 2013. 36

[57] S. Workman, R. Souvenir, and N. Jacobs, "Scene shape estimation from multiple partly cloudy days," *Computer Vision and Image Understanding*, vol. 134, pp. 116–129, 2015, image Understanding for Real-world Distributed Video Networks. 36

[58] H. Kurihata, T. Takahashi, I. Ide, Y. Mekada, H. Murase, Y. Tamatsu, and T. Miyahara, "Rainy weather recognition from in-vehicle camera images for driver assistance," in *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, 2005, pp. 205–210. 36

[59] L. Shen and P. Tan, "Photometric stereo and weather estimation using internet images," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1850–1857. 36

[60] C. Lu, D. Lin, J. Jia, and C.-K. Tang, "Two-class weather classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2510–2524, 2017. 36, 37, 43, 48, 54

[61] Z. Chen, F. Yang, A. Lindner, G. Barrenetxea, and M. Vetterli, "Howis the weather: Automatic inference from images," in *2012 19th IEEE International Conference on Image Processing*, 2012, pp. 1853–1856. 36

[62] H. Song, Y. Chen, and Y. Gao, "Weather condition recognition based on feature extraction and k-nn," *Advances in Intelligent Systems and Computing*, vol. 215, pp. 199–210, 09 2014. 36

[63] J. An, Y. Chen, and H. Shin, "Weather classification using convolutional neural networks," in *2018 International SoC Design Conference (ISOCC)*, 2018, pp. 245–246. 36

[64] S. Narasimhan and S. Nayar, "Shedding light on the weather," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1, 2003, pp. I–I. 36

[65] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Trans. Graph.*, vol. 33, pp. 149:1–149:11, 07 2014. 37

[66] Z. Zhang and H. Ma, "Multi-class weather classification on single images," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 4396–4400. 37, 48

[67] M. Elhoseiny, S. Huang, and A. Elgammal, "Weather classification with deep convolutional neural networks," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 3349–3353. 37

[68] C. Ancuti, C. O. Ancuti, and C. De Vleeschouwer, "D-hazy: A dataset to evaluate quantitatively dehazing algorithms," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2226–2230. 38

[69] H. Bhandari, S. Palit, S. Chowdhury, and P. Dey, "Can a camera tell the weather?" in *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2021, pp. 1–6. 43

[70] I. I. A. Groen, S. Ghebreab, H. Prins, V. A. F. Lamme, and H. S. Scholte, "From image statistics to scene gist: Evoked neural activity reveals transition from low-level natural image structure to scene category," *The Journal of Neuroscience*, vol. 33, pp. 18 814 – 18 824, 2013. 47

[71] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002. 47

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 48

[73] Y.-F. Liu, D. Jaw, w.-c. Yeh, and J.-N. Hwang, "Desnownet: Context-aware deep network for snow removal," *IEEE Transactions on Image Processing*, vol. PP, 08 2017. 48

[74] Z. Li, Y. Li, J. Zhong, and Y. Chen, "Multi-class weather classification based on multi-feature weighted fusion method," *IOP Conference Series: Earth and Environmental Science*, vol. 558, no. 4, p. 042038, aug 2020. 53, 54

[75] K. Gu, J. Qiao, and X. Li, "Highly efficient picture-based prediction of pm2.5 concentration," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3176–3184, 2019. 56, 63, 64

[76] C. Liu, F. Tsow, Y. Zou, and N. Tao, "Particle pollution estimation based on image analysis," *PLOS ONE*, vol. 11, no. 2, pp. 1–14, 02 2016. 56, 57, 64, 70, 71

[77] C.-H. Hsieh, K.-Y. Chen, M.-Y. Jiang, J.-J. Liaw, and J. Shin, "Estimation of pm2.5 concentration based on support vector regression with improved dark channel prior and high frequency information in images," *IEEE Access*, vol. 10, pp. 48 486–48 498, 2022. 56, 72

[78] A. Chakma, B. Vizena, T. Cao, J. Lin, and J. Zhang, "Image-based air quality analysis using deep convolutional neural network," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3949–3952. 56, 73

[79] N. Nilesh, I. Patwardhan, J. Narang, and S. Chaudhari, "Iot-based aqi estimation using image processing and learning methods," in *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*, 2022, pp. 1–5. 56

[80] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3522–3533, 2015. 63

[81] Y. Han, M. Qi, Y. Chen, H. Shen, J. Liu, Y. Huang, H. Chen, W. Liu, X. Wang, J. Liu, B. Xing, and S. Tao, "Influences of ambient air pm2.5 concentration and meteorological condition on the indoor pm2.5 concentrations in a residential apartment in beijing using a new approach," *Environmental pollution (Barking, Essex : 1987)*, vol. 205, pp. 307–314, 06 2015. 69

[82] Z. Zhang, H. Ma, H. Fu, L. Liu, and C. Zhang, "Outdoor air quality level inference via surveillance cameras," *Mobile Information Systems*, vol. 2016, p. 9825820, Jun 2016. 70, 71

[83] A. Schatz, G. Thurston, R. Isaacs, and R. Husar, "Relationships between aerosol extinction coefficients derived from airport visual range observations and alternative measures of airborne particle mass," *Journal of the Air Pollution Control Association*, vol. 35, 11 1985. 71

[84] N. Hyslop, "Impaired visibility: the air pollution people see," *Atmospheric Environment*, vol. 43, pp. 182–195, 01 2009. 71

[85] B. Yang and Q. Chen, "Pm2.5 concentration estimation based on image quality assessment," in *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, 2017, pp. 676–681. 72

[86] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1956–1963. 72

[87] X. Liu, Z. Song, E. Ngai, J. Ma, and W. Wang, "Pm2:5 monitoring using images from smartphones in participatory sensing," in *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2015, pp. 630–635. 72

[88] G. Yue, K. Gu, and J. Qiao, "Effective and efficient photo-based pm2.5 concentration estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 10, pp. 3962–3971, 2019. 72

[89] Z. Chen, T. Zhang, Z. Chen, Y. Xiang, Q. Xuan, and R. P. Dick, "Hvaq: A high-resolution vision-based air quality dataset," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021. 72

[90] Y. Li, J. Huang, and J. Luo, "Using user generated online photos to estimate and monitor air pollution in major cities," in *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, ser. ICIMCS '15. New York, NY, USA: Association for Computing Machinery, 2015. 72

[91] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2015. 72

[92] C. Zhang, J. Yan, C. Li, H. Wu, and R. Bie, "End-to-end learning for image-based air quality level estimation," *Machine Vision and Applications*, vol. 29, 05 2018. 72

[93] A. Chakma, B. Vizena, T. Cao, J. Lin, and J. Zhang, "Image-based air quality analysis using deep convolutional neural network," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3949–3952. 72

[94] J. Ma, K. Li, Y. Han, and J. Yang, "Image-based air pollution estimation using hybrid convolutional neural network," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 471–476. 72

[95] Z. Wang, W. Zheng, C. Song, Z. Zhang, J. Lian, S. Yue, and S. Ji, "Air quality measurement based on double-channel convolutional neural network ensemble learning," *IEEE Access*, vol. 7, pp. 145 067–145 081, 2019. 73

[96] L. Liu, W. Liu, Y. Zheng, H. Ma, and C. Zhang, "Third-eye: A mobilephone-enabled crowdsensing system for air quality monitoring," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, mar 2018. 73

[97] K. Li, J. Ma, H. Li, Y. Han, X. Yue, Z. Chen, and J. Yang, "Discern depth under foul weather: Estimate $PM_{2.5}$ for depth inference," *IEEE Transactions on Industrial Informatics*, vol. PP, pp. 1–1, 09 2019. 73

[98] S. Song, J. C. K. Lam, Y. Han, and V. O. K. Li, "Resnet-lstm for real-time pm2.5 and pm10 estimation using sequential smartphone images," *IEEE Access*, vol. 8, pp. 220 069–220 082, 2020. 73

[99] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2366–2374. 73, 78, 84

[100] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011, 2018. 73

[101] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth International Conference on 3D Vision (3DV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2016, pp. 239–248. 73

[102] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Transactions on Pattern Analysis &; Machine Intelligence*, vol. 38, no. 10, pp. 2024–2039, oct 2016. 73

[103] Z. Yu, L. Jin, and S. Gao, "P^2net: Patch-match and plane-regularization for unsupervised indoor depth estimation," in *ECCV*, 2020. 73

[104] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep Ordinal Regression Network for Monocular Depth Estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 73

[105] S. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," 06 2021, pp. 4008–4017. 74

[106] K. Parida, S. Srivastava, and G. Sharma, "Beyond image to depth: Improving depth prediction using echoes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 74

[107] T. Leistner, R. Mackowiak, L. Ardizzone, U. Kuthe, and C. Rother, "Towards multimodal depth estimation from light fields," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2022, pp. 12 943–12 951. 74

[108] R. Garg, V. K. B.G., G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 740–756. 74

[109] M. Poggi, F. Tosi, and S. Mattoccia, "Learning monocular depth estimation with unsupervised trinocular assumptions," in *6th International Conference on 3D Vision (3DV)*, 2018. 74

[110] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017. 74

[111] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth prediction," October 2019. 74

[112] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *The International Conference on Computer Vision (ICCV)*, October 2019. 74

[113] R. Peng, R. Wang, Y. Lai, L. Tang, and Y. Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 74

[114] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. Madison, WI, USA: Omnipress, 2010, p. 807–814. 75

[115] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1511.07122, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:17127188 75

[116] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 81

[117] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 3–19. 82

[118] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594692 83

[119] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 707–724. 84

[120] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. 92, 95

[121] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 2261–2269. 92, 95

[122] 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*, ser. AAAI'17. AAAI Press, 2017, p. 4278–4284. 92, 95

[123] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. 92, 95

[124] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. 95

[125] P. Drews Jr, E. do Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *2013 IEEE International Conference on Computer Vision Workshops*, 2013, pp. 825–830. 104

[126] Y. Wang, H. Liu, and L.-P. Chau, "Single underwater image restoration using adaptive attenuation-curve prior," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 3, pp. 992–1002, 2018. 104

[127] C. Li, J. Guo, S. Chen, Y. Tang, Y. Pang, and J. Wang, "Underwater image restoration based on minimum information loss principle and optical properties of underwater imaging," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1993–1997. 104

[128] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1682–1691. 104

[129] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, and C. Li, "Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement," *IEEE Transactions on Image Processing*, vol. 31, pp. 3997–4010, 2022. 104, 114

[130] P. Zhuang, J. Wu, F. Porikli, and C. Li, "Underwater image enhancement with hyper-laplacian reflectance priors," *IEEE Transactions on Image Processing*, vol. 31, pp. 5442–5455, 2022. 104, 114

[131] M. S. Hitam, E. A. Awalludin, W. N. Jawahir Hj Wan Yussof, and Z. Bachok, "Mixture contrast limited adaptive histogram equalization for underwater image enhancement," in *2013 International Conference on Computer Applications Technology (ICCAT)*, 2013, pp. 1–5. 105

[132] K. Iqbal, M. Odetayo, A. James, R. A. Salam, and A. Z. H. Talib, "Enhancing the low quality images using unsupervised colour correction method," in *2010 IEEE International Conference on Systems, Man and Cybernetics*, 2010, pp. 1703–1709. 105

[133] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 81–88. 105

[134] X. Li, G. Hou, L. Tan, and W. Liu, "A hybrid framework for underwater image enhancement," *IEEE Access*, vol. 8, pp. 197 448–197 462, 2020. 105

[135] J. Li, K. A. Skinner, R. Eustice, and M. Johnson-Roberson, "Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation Letters (RA-L)*, 2017, accepted. 105

[136] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE Transactions on Image Processing*, vol. 30, pp. 4985–5000, 2021. 105, 112, 114

[137] Z. Fu, W. Wang, Y. Huang, X. Ding, and K.-K. Ma, "Uncertainty inspired underwater image enhancement," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 465–482. 105, 114

[138] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE Transactions on Image Processing*, vol. 32, pp. 3066–3079, 2023. 105, 114

[139] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016. 107

[140] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4861–4875, 2020. 112