Addressing class imbalance problems to improve animal detection through aerial image data

A dissertation submitted in

partial fulfilment for the degree of

Master of Technology

Computer Science

by

Suryang Koushal

Roll No. CS2332

under the supervision of

Dr. Sarbani Palit

Dr. Ujjwal Verma

(Manipal Institute of Technology)

Indian Statistical Institute, Kolkata

June, 2024

Certificate

This is to certify that the dissertation titled "Addressing Class Imbalance Problems to Improve Animal Detection through Aerial Image Data" submitted by Suryang Koushal to the Indian Statistical Institute, Kolkata, in partial fulfilment of the requirements for thee MTech in Computer Science, is a original piece of research conducted under my supervision and guidance. I hereby confirm that this dissertation adheres to the academic and administrative norms as set by the Institute.

Sarbani Palit 11.6.2025

Dr. Sarbani Palit
CVPR Unit
Indian Statistical Institute
Kolkata - 700108, India

Acknowledgement

I would like to sincerely thank Dr. Sarbani Palit, my advisor at the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, for her invaluable guidance, steady encouragement, and continuous support throughout the course of this research.Her guidance has significantly influenced the scope and substance of this research.

I am also deeply grateful to Dr. Ujjwal Verma for his insightful guidance and thoughtful suggestions, which have significantly enhanced the quality and clarity of this dissertation.

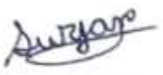
A special note of thanks goes to Harsh Bhandari, Senior Research Fellow at ISI, whose essential support, technical inputs, and consistent feedback were crucial to the development and refinement of this project.

I would like to express my appreciation to all the faculty members at the Indian Statistical Institute for their academic guidance and the solid foundation they provided in the formative stages of my research. I am particularly thankful to Anurag Pal for his dedicated mentorship and unwavering support.

I am also thankful to my friends for their continued motivation and readiness to offer help whenever needed. Lastly, I extend heartfelt thanks to everyone who has contributed to this journey in any way—your support, even if not acknowledged by name, has been truly significant and appreciated.

Declaration

I, Suryang Koushal, with Roll No. CS2332, hereby declare that the material presented in the dissertation titled "Addressing Class Imbalance Problems to Improve Animal Detection through Aerial Image Data" represents authentic work completed by me for Master of Technology in Computer Science at the Indian Statistical Institute, Kolkata. I confirm that no section of this document has been copied or reproduced from other sources without appropriate acknowledgment. I recognize that any violation involving uncredited content will lead to serious academic consequences.



Suryang Koushal
M.Tech (CS), CS2332
Indian Statistical Institute

Abstract

Monitoring animal populations in wildlife reserves is essential for conservation, especially for endangered species, but manual censuses are costly, risky, and logistically challenging due to vast, inaccessible terrains. Unmanned Aerial Vehicles (UAVs) with digital cameras provide a safer, scalable solution for collecting aerial imagery to estimate animal populations. However, semi-automated processing of these images faces significant challenges due to class imbalance in datasets, including foreground-background disparities, where background terrain dominates over sparse animal instances, and inter-class imbalances from uneven species representation and varied visual appearances (e.g., species, sizes, fur patterns) against diverse backgrounds like deserts or forests. These imbalances hinder Convolutional Neural Networks (CNNs) used for object detection, leading to inaccurate population estimates. This project addresses these issues using a dataset of 561 aerial images from Tsavo National Parks (March 2014) and Laikipia-Samburu Ecosystem (May 2015), collected by the Kenya Wildlife Service. We propose a clustering-based approach to categorize background terrain into distinct classes (e.g., desert, grassland), aiming to mitigate imbalances and improve animal detection accuracy in UAV imagery, supporting reliable, data-driven conservation strategies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Monitoring wildlife numbers in expansive conservation zones is crucial for informed preservation efforts, particularly for tracking endangered species. Traditional methods, such as manual ground-based surveys, are labor-intensive, expensive, and sometimes dangerous. In contrast, Unmanned Aerial Vehicles (UAVs) equipped with digital cameras have emerged as a reliable and cost-effective alternative, enabling high-resolution aerial imagery to be collected safely and efficiently.

Convolutional Neural Networks (CNNs) are widely used for processing UAV images and detecting animals.Despite their utility, these models tend to underperform in scenarios where class frequencies are imbalanced and animals appear infrequently and in small regions. This imbalance skews the model's learning, leading to poor generalization and missed detections.

To tackle this challenge, we utilized RetinaNet — a single-stage detection model recognized for its use of focal loss mechanism that down-weights easy background examples and emphasizes harder, minority-class instances. We further improved its performance by dividing the background into several semantic sub-classes, aiming to reduce the dominance of a single generic background label.

In addition, we explored DETR (DEtection TRansformer), a transformer-based object detection model that leverages global self-attention. Unlike anchor-based CNNs, DETR can model long-range dependencies and attend to small, sparse objects in cluttered scenes. This makes it particularly suitable for imbalanced datasets like ours, where foreground-background distribution is heavily skewed.

Chapter 2

Related Work

Recent advancements in deep learning have significantly enhanced wildlife detection capabilities in aerial imagery. A core challenge in this domain is the pronounced *class imbalance* between foreground animals and the vast background, which often leads to decreased performance in detecting small or rare species.

Kellenberger et al. [7] addressed this issue by outlining best practices for dealing with heavily imbalanced UAV datasets. Their strategies included data balancing, customized sampling, and loss function tuning to improve rare animal detection. Building upon this, Zheng et al. [13] proposed a combination of *self-supervised pretraining* and *controlled data augmentation* to improve recognition of rare species, especially in settings with limited annotated data.

Comprehensive reviews by Oksuz et al. [10] and Chen et al. [4] explored various imbalance scenarios—such as scale, spatial distribution, and semantic class imbalance—and recommended techniques like focal loss and intelligent sampling to mitigate their effects. These approaches are directly applicable to modern detectors such as RetinaNet [8] and Faster R-CNN [11], which incorporate such mechanisms to improve detection robustness.

The use of pretrained models on remote sensing datasets has also shown significant promise. Open-source models trained on the AID dataset [**?**] have proven effective in extracting high-level features from aerial imagery, facilitating downstream tasks such as clustering or transfer learning.

Eikelboom et al. [5] demonstrated that deep learning methods not only matched but exceeded human performance in animal counting from aerial images. Their findings showed that automated detection led to improved statistical precision and substantial cost reductions in wildlife monitoring efforts.

To better capture background class variability, Sarkar and Ghosh [12] developed clustering strategies suited for scenarios involving high-dimensional features and few observations. Their use of Jump Statistics guided our strategy to cluster images into terrain-aware background categories, enriching the model's understanding of visual context and reducing confusion between background and animal classes.

Together, these works provide a strong basis for addressing the challenges posed by class imbalance. integrating domain-specific priors, and enhancing the effectiveness of object detection systems in ecological monitoring using UAV imagery.

Chapter 3

Dataset

3.0.1 About

The dataset used in this study was provided by the Kenya Wildlife Service. It comprises aerial images captured during animal surveys using aircraft-mounted cameras. These surveys were conducted in two wildlife regions: the Tsavo National Parks (March 2014) and the Laikipia-Samburu Ecosystem (May 2015), both located in Kenya.

3.0.2 Dataset Summary

- **Total number of images:** 561
- **Data split:**
  - Training set: 393 images (70%)
  - Validation set: 56 images (10%)
  - Test set: 112 images (20%)
- **Image dimensions:** Each image is either 4603 × 3068 or 5184 × 3452 pixels
- **Original annotations:**
  - Elephants: 1,319 instances
  - Giraffes: 1,109 instances

– Zebras: 1,877 instances

3.0.3 Preprocessing

RetinaNet

- Due to high image resolution and limited memory, each training image was divided into 42 tiles with 200-pixel overlaps.
- Tile dimensions averaged approximately 900 × 700 pixels.
- Overlaps ensured that animals crossing tile boundaries were fully visible in at least one tile.
- Only fully visible animals were retained; partial animals were cropped out and treated as background.
- Animals not belonging to the target classes (elephant, giraffe, zebra) were also considered background.
- Data augmentation included horizontal flipping.
- Only tiles containing at least one fully visible animal were used in training (roughly 10% of generated tiles).

DETR

- Similar memory issues led to splitting each image into 16 tiles.
- Tile sizes were either 1150 × 767 or 1296 × 863 pixels.
- Extra augmentation techniques were employed to enhance the model's generalization capabilities.
- Adjacent tiles within rows were combined with 50% horizontal overlap to create additional samples.
- These overlapping tiles were added to the training dataset.

Table 3.1: Counts of images and animal occurrences for each species within the train, validation, and test datasets.

Set	**Images**	**Elephant**	**Giraffe**	**Zebra**
Training	393	2,640	2,160	4,182
Validation	56	140	93	219
Test	112	288	261	301

Figure 3.1: An example of image set with bounding and classes marked

Figure 3.2: A patch from Figure 3.1 containing some elephants

Chapter 4

Methodology

Introduction to Object Detection

Object detection refers to the task of pinpointing and classifying multiple items of varying shapes, types, and scales within an image. This process typically involves three fundamental steps:

1. Identify all object instances present in the scene.
2. Enclose each identified object using a bounding box.
3. Provide the class label and bounding box coordinates, either as the center with width and height or as corner points.

Single-Stage Object Detectors

- These models instantly predict object classes and bounding boxes from input images in one pass.
- Feature extraction is carried out using deep convolutional layers with filters to capture spatial hierarchies.
- Well-suited for applications requiring quick inference times.
- **Common Models:** SSD, YOLO, RetinaNet.

Transformer-Based Detectors

- These treat detection as a direct set prediction task, leveraging attention mechanisms.
- Employ encoder-decoder transformers to grasp global scene relationships and context.
- Remove reliance on anchor boxes and region proposal networks.
- Generally slower than CNN-based detectors but more effective in cluttered or imbalanced environments.
- **Notable Architectures:** DETR, Deformable DETR, DINO.

Focal Loss

Focal Loss is designed specifically to reduce the impact of class imbalance during training by placing more emphasis on challenging examples during training. Its formulation is given by:

$$\mathrm{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \tag{4.1}$$

Here:

- p_t indicates the probability the model assigns to the true category.
- γ is a tunable parameter that adjusts the degree to which easier samples are down-weighted.

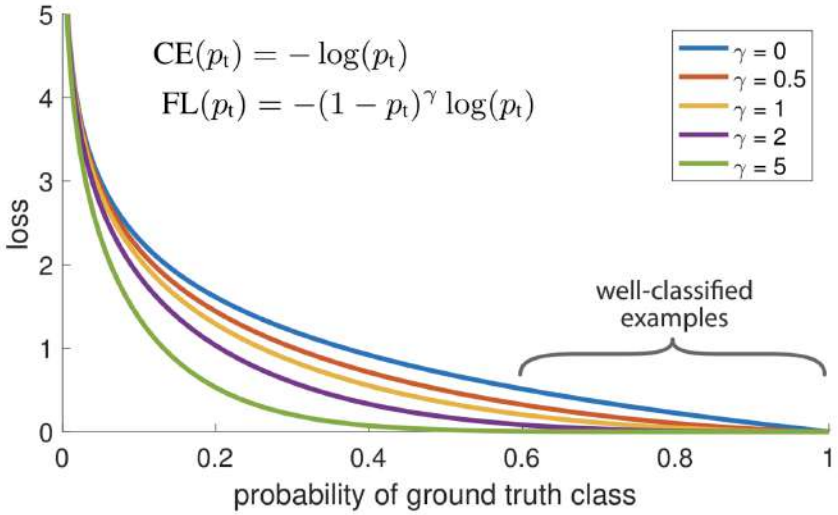


Figure 4.1: Focal loss

The component $(1 - p_t)^{\gamma}$ serves as a scaling factor that minimizes the influence of confidently predicted examples (when p_t is near 1), thereby directing the model's attention toward instances that are more difficult to classify correctly.

4.0.1 Models

RetinaNet

RetinaNet is a one-stage object detection model built from several fundamental components:

- **Bottom-Up Feature Pathway:** Utilizes a convolutional backbone (such as ResNet) to generate hierarchical feature maps at multiple scales.
- **Top-Down Architecture with Lateral Links:** Enhances resolution by upsampling coarse features and combining them with higher-resolution counterparts.
- **Classification Head:** Outputs class probabilities for each anchor across all spatial positions.
- **Regression Head:** Outputs offsets for adjusting anchor boxes to better match detected objects.

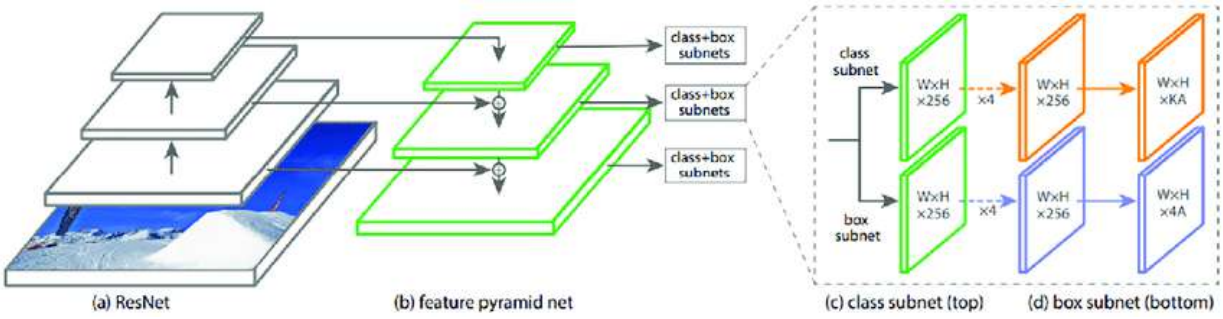


Figure 4.2: Structure of the RetinaNet architecture

Clustering Background Classes

- Visual analysis reveals that the dataset comprises diverse landscapes such as forests, deserts, and grasslands—terrains that often exhibit similar visual traits.
- While manually labeling terrain types with expert input is feasible, it requires considerable time and resources.

- To overcome this, we assume that each image contains a dominant background type. Images are then clustered, and the cluster index is assigned as the label for background elements.

High-Dimensional Clustering Strategy

- Basic clustering algorithms like K-means struggle in high-dimensional feature spaces. To address this, a variant based on Mean Absolute Deviation Distance (MADD) is employed.
- Due to the large size of each image, direct clustering is impractical. Instead, we extract abstract representations using pretrained networks.
- These networks offer a form of transfer learning, allowing image features to be compacted into a semantically rich vector representation.
- A ResNet-34 model trained on the AID (Aerial Image Dataset) is used for this purpose, which has exposure to satellite images across various land cover classes.
- We discard the final fully connected layer and use the activations from the last convolutional block as descriptors.
- To determine the best number of clusters, we rely on Jump Statistics—well-suited for high-dimensional data. This method suggests $k = 9$ as the optimal choice.

Visual Inspection of Clusters

figure 4.3 highlights samples grouped into a single cluster, showcasing the model's ability to separate similar visual backgrounds.

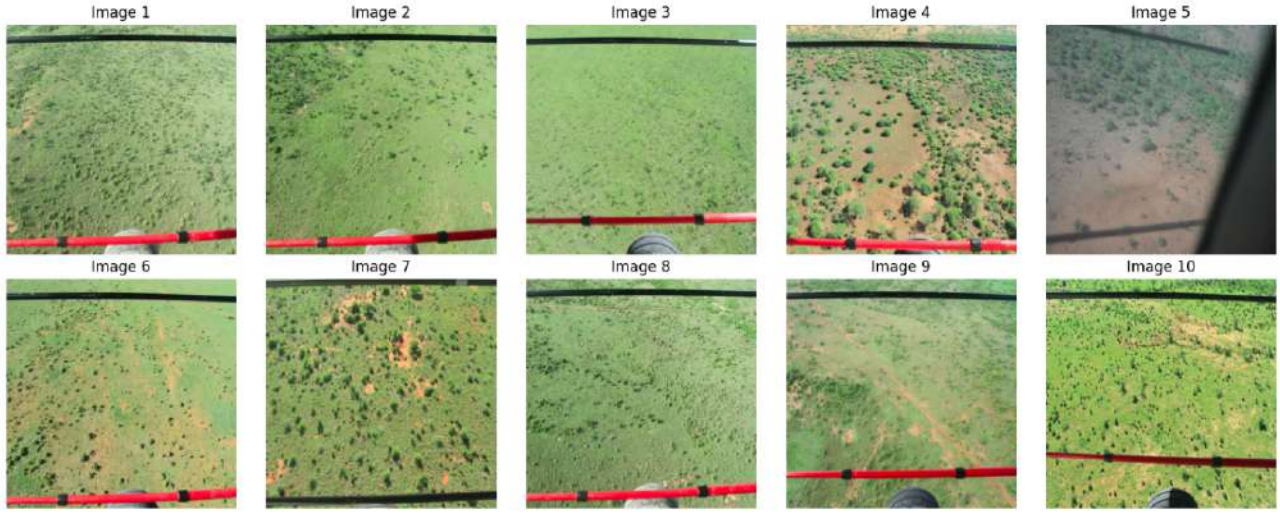


Figure 4.3: Instances of grassland regions classified within a single cluster.

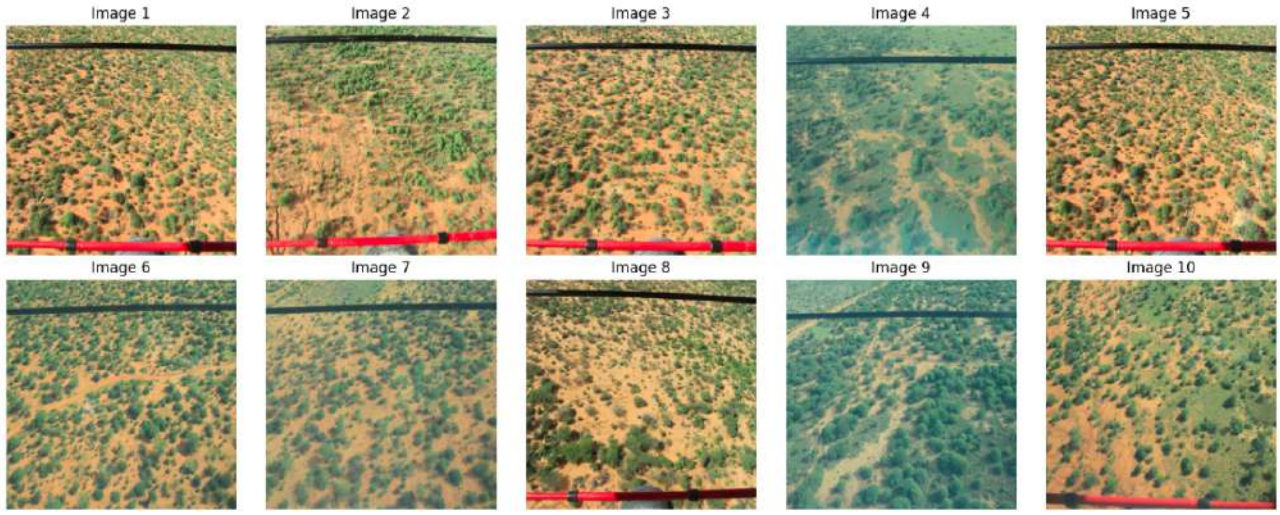


Figure 4.4: Desert regions with similar vegetation clustered together

DETR (DEtection TRansformer)

DETR introduces a novel approach to object detection by formulating it as a direct set prediction task. It does away with hand-crafted components like anchors and region proposals.

- **Feature Extractor:** A backbone CNN such as ResNet is used to convert raw images into compact feature maps.
- **Transformer Module:**
  - The encoder applies multi-head self-attention to understand global dependencies across the image.
  - The decoder operates on fixed object queries, producing predictions for potential objects in the scene.
- **Set Prediction Head:** Produces a constant-size list of detection outputs, each linked to an object or marked as background.
- **Assignment via Hungarian Algorithm:** Ensures a unique match between predictions and ground truth using optimal bipartite matching.
- **Strengths:** DETR simplifies the overall detection pipeline and performs particularly well in crowded scenes or when class distribution is skewed.

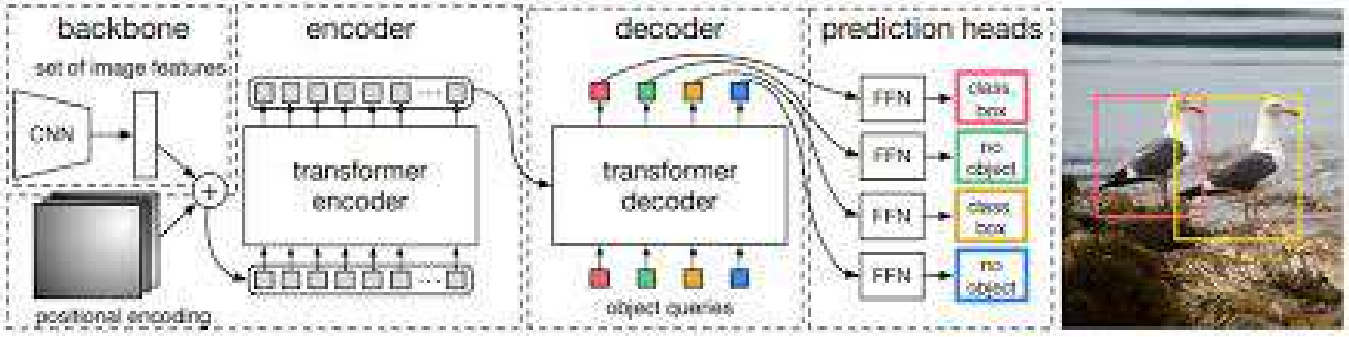


Figure 4.5: Overview of the DETR detection framework

4.0.2 Training

- All models were implemented using PyTorch and trained with GPU acceleration enabled via CUDA.
- The training was conducted on a system equipped with an NVIDIA Tesla P6 GPU and an Intel Xeon Platinum 8164 CPU, running Ubuntu (x86_64).
- RetinaNet training was limited to a maximum of 40 epochs or a duration of 10 hours, depending on which occurred first.
- Both Adam and SGD optimizers were explored for RetinaNet, each tested with different learning rate configurations.
- DETR training was carried out for up to 150 epochs, incorporating early stopping based on stagnation in training loss.
- Early stopping was configured with a patience of 10 epochs, halting training if no improvement in the training loss was detected.
- Gradient accumulation was employed every 4 iterations to reduce memory overhead during large batch training.
- Mixed precision training via `torch.cuda.amp` was initially set up, but remained disabled due to CPU-only fallback scenarios.
- For DETR, the AdamW optimizer was chosen, with a learning rate of 1×10^{-6} and a weight decay rate of 5×10^{-4}.
- Model evaluation was performed after each epoch using validation loss to monitor generalization capability.
- Data preprocessing and annotation conversion were managed using the `DetrImageProcessor` utility from the HuggingFace Transformers library.
- Trained model weights and the corresponding processor settings were saved using HuggingFace's `save_pretrained` method for reproducibility.
- Learning rates were tuned by observing trends in the validation mAP (mean Average Precision).

- The checkpoints with the best validation mean Average Precision were retained as the optimal models.

4.0.3 Performance Evaluation

Intersection over Union (IoU)

Intersection over Union (IoU) is a core evaluation metric that quantifies the agreement between predicted and actual bounding boxes or segmentation masks. It quantifies the degree of spatial overlap between predicted and actual object regions.

The IoU is computed using the following formula:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{4.2}$$

Where:

- A refers to the region predicted by the model,
- B corresponds to the actual (ground truth) region,
- $|A \cap B|$ is the intersection area of A and B,
- $|A \cup B|$ represents the union area of A and B.

IoU values lie between 0 and 1:

- An IoU value of 1 indicates a complete match between the predicted output and the actual annotation,
- An IoU value of 0 signifies a total lack of overlap between the two regions.

This metric plays a crucial role in both training and evaluation phases, as it provides a measure of how precisely the model identifies and localizes target objects.

Mean Average Precision (mAP)

Mean Average Precision (mAP) is a widely used metric for evaluating object detection performance, incorporating both precision and recall over various classes and confidence thresholds.

Precision and Recall

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.4}$$

Where:

- TP = True Positive detections,
- FP = False Positive detections,
- FN = False Negatives (missed objects).

Average Precision (AP)

AP represents the area under the Precision-Recall curve and is computed as:

$$\text{AP} = \int_0^1 P(R)\, dR \tag{4.5}$$

It reflects how well precision is maintained over increasing levels of recall for a specific class.

mAP@0.5

This is the mean of Average Precision scores at an IoU threshold of 0.5 across all classes:

$$\text{mAP@0.5} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_{0.5}^{(i)} \tag{4.6}$$

Where:

- N is the total number of target classes,
- $\text{AP}_{0.5}^{(i)}$ is the AP for class i using IoU threshold of 0.5.

mAP@0.5:0.95

To evaluate detector robustness across multiple overlap thresholds, mAP is averaged over a range of IoU thresholds from 0.5 to 0.95 (in steps of 0.05):

$$\text{mAP@0.5:0.95} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{11} \sum_{t=0.5}^{0.95} \text{AP}_{t}^{(i)} \right) \tag{4.7}$$

Where:

- t spans IoU values from 0.5 to 0.95 in increments of 0.05,
- $\text{AP}_{t}^{(i)}$ is the AP for class i at threshold t.

Both mAP@0.5 and mAP@0.5:0.95 are widely used in benchmarking challenges, with the latter offering a more nuanced and rigorous evaluation of detection performance across varying overlap conditions.

Chapter 5

Experiments and Results

This chapter summarizes the experimental findings from the object detection models trained on satellite imagery. The evaluation focuses on standard performance metrics such as mAP@50 and mAP@50:95 on the test set. Each model is described with its training setup and the results achieved by the best-performing configuration.

5.1 RetinaNet

5.1.1 Overview

Training Configuration:

- `Optimizer`: Adam
- `Learning Rate`: 1e-5

Performance of Best Model:

- `mAP@50 (Test)`: 76.88%
- `mAP@50:95 (Test)`: 43.10%
- `Epoch`: 33

5.2 RetinaNet with 9 Background Classes

5.2.1 Overview

Training Configuration:

- `Optimizer:` Adam
- `Learning Rate:` 1e-5

Performance of Best Model:

- `mAP@50 (Test):` 78.21%
- `mAP@50:95 (Test):` 44.93%
- `Epoch:` 35

5.3 DETR with Focal Loss

5.3.1 Overview

Training Configuration:

- `Optimizer:` AdamW (Adam with decoupled weight decay)
- `Learning Rate:` 1e-6
- `Weight Decay:` 5e-4

Performance of Best Model:

- `mAP@50 (Test):` 71.80%
- `mAP@50:95 (Test):` 41.24%
- `Epoch:` 135

Chapter 6

Predictions

6.1 From Retinanet

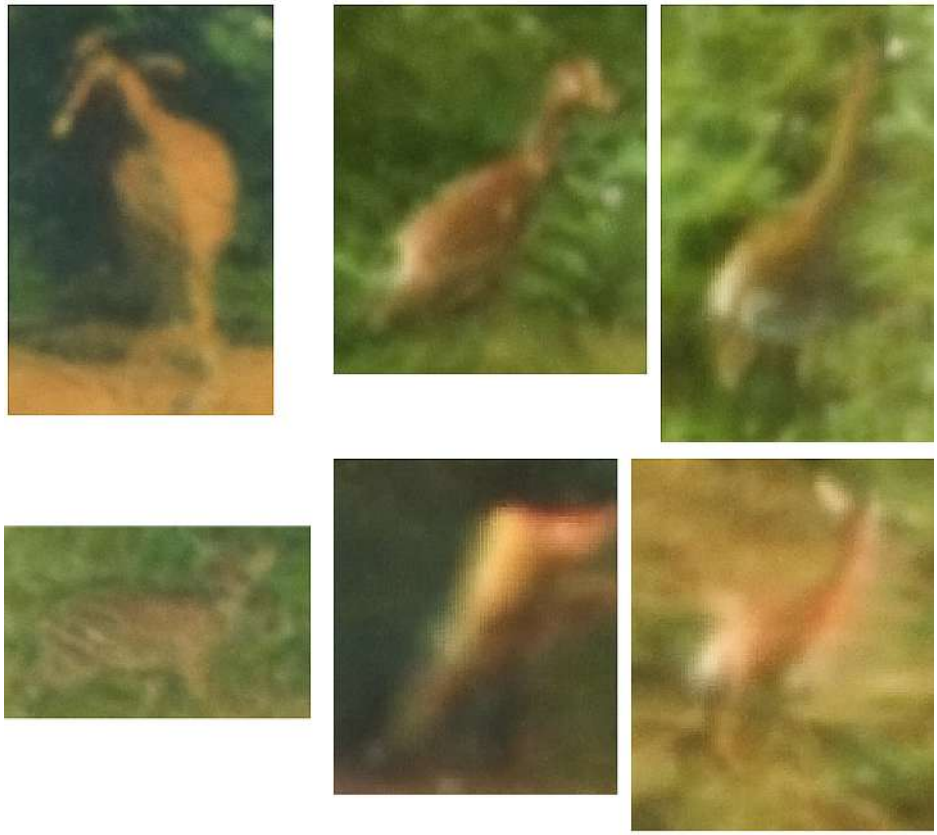


Figure 6.1: Examples of False positive with high confidence score (predicted as Giraffe)

Figure 6.2: Examples of annotated image of Giraffe which model considered background

6.2 From Retinanet with 9 bg classes

Figure 6.3: Examples of False positive with high confidence score (predicted as Giraffe)

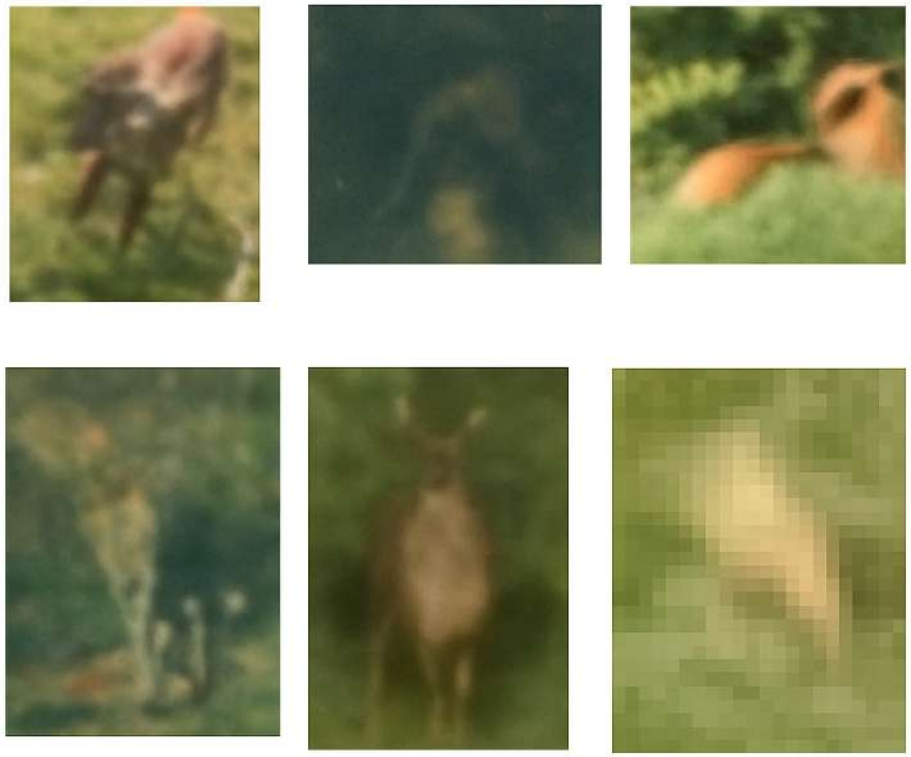


Figure 6.4: Examples of annotated image of Giraffe which model considered background

6.3 From DETR with focal loss

Figure 6.5: Examples of False positive with high confidence score (predicted as Giraffe)

Figure 6.6: Examples of annotated image of Giraffe which model considered background

Chapter 7

Conclusion and Future Work

This study aimed to address the class imbalance challenges prevalent in animal detection using aerial imagery, particularly from UAVs. Traditional CNN-based object detection models often struggle with performance degradation due to the disproportionate presence of background pixels compared to the relatively small and sparse occurrences of animal classes. To mitigate this, we evaluated RetinaNet, a one-stage detector known for its robustness against imbalance through the use of focal loss, and DETR, a transformer-based detector capable of capturing long-range dependencies and modeling sparse foregrounds effectively.

To further reduce background dominance, we introduced a novel background clustering technique. By using feature extraction from a pretrained ResNet-34 model trained on aerial imagery, followed by high-dimensional clustering (using a MADD variant of k-means), we segmented background terrain into nine distinct categories. This significantly improved foreground detection accuracy by reducing background misclassification and allowing the detector to distinguish between semantically meaningful background classes.

Experimental results confirmed that both RetinaNet with clustered backgrounds and DETR with focal loss achieved improved detection metrics (mAP@50 and mAP@50:95) over the standard configurations. In particular, the RetinaNet model with background clustering showed the best balance between precision and generalization.

Future Work.

While this study addressed foreground-background imbalance effectively, several promising directions remain open for exploration. First, integrating temporal consistency across video sequences from UAVs could improve object tracking and detection stability. Second, self-supervised pretraining or contrastive learning methods could be incorporated to further improve feature representations under limited labels.

Finally, real-time deployment constraints such as computational efficiency, energy usage on UAVs, and adaptive model compression are essential considerations for large-scale ecological applications.

Bibliography

[1] Remote sensing pretrained models. `https://github.com/lsh1994/remote-sensing-pretrained-models`. Accessed: 2025-06-10.

[2] M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[4] Joya Chen, Qi Wu, Dong Liu, and Tong Xu. Foreground-background imbalance problem in deep object detectors: A review. In *IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2020.

[5] Jasper A. J. Eikelboom, Johan Wind, Eline van de Ven, Lekishon M. Kenana, Bradley Schroder, Henrik J. de Knegt, Frank van Langevelde, and Herbert H. T. Prins. Improving the precision and accuracy of animal population estimates with aerial image object detection. *Methods in Ecology and Evolution*, 10(9):1415–1425, 2019.

[6] Fei Huang and Yonghong Zhang. A review of image retrieval methods for the internet. *International Journal of Computer Science and Network Security (IJCSNS)*, 9(2):23–30, 2009.

[7] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. Detecting mammals in uav images: Best practices to address a substantially imbalanced dataset with deep learning. *CoRR*, abs/1806.11368, 2018.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[10] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3388–3415, 2021.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[12] Soham Sarkar and Anil K. Ghosh. On perfect clustering of high dimension, low sample size data. *arXiv preprint arXiv:1612.09121*, 2016.

[13] X. Zheng, B. Kellenberger, R. Gong, I. Hajnsek, and D. Tuia. Self-supervised pretraining and controlled augmentation improve rare wildlife recognition in uav images. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 732–741, Montreal, BC, Canada, 2021. IEEE.