SkyFusion: Aerial Object Detection

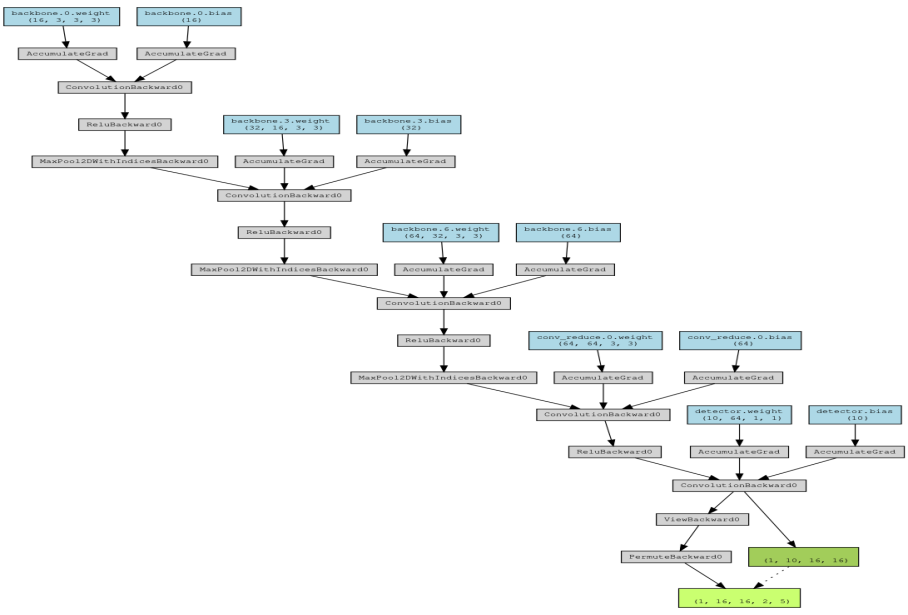
ADITYA RAJ

February 2025

1 Introduction

The primary objective was to develop a custom small object detection model trained from scratch—without leveraging any pre-trained models—and to evaluate its performance based on the mean Average Precision (mAP) metric. The study utilizes the SkyFusion: Aerial Object Detection dataset, which comprises satellite images featuring three object classes: *Airplane*, *Vehicle*, and *Ships*. The dataset employed in this study represents only 10% of the original SkyFusion data, removing the need for additional downsampling. For benchmarking, we compare our custom model against Faster R-CNN framework with a ResNet-50-FPN backbone from the PyTorch library to benchmark the custom built model.

2 Methodology



The SimpleDetector adopts a streamlined, one-stage grid-based approach. It employs a custom convolutional neural network comprising three sequential convolutional blocks with ReLU activations and max pooling to extract features from the input image. The network divides the image into a fixed grid, where each cell predicts a fixed number of bounding boxes with five parameters each (i.e., x, y, w, h, and confidence).

SimpleDetector employs a custom loss function that weighs coordinate and confidence losses differently based on the presence of an object. Its inference process involves a single forward pass

to generate grid predictions, which are then converted into absolute bounding box coordinates and filtered using a custom non-maximum suppression (NMS) method.

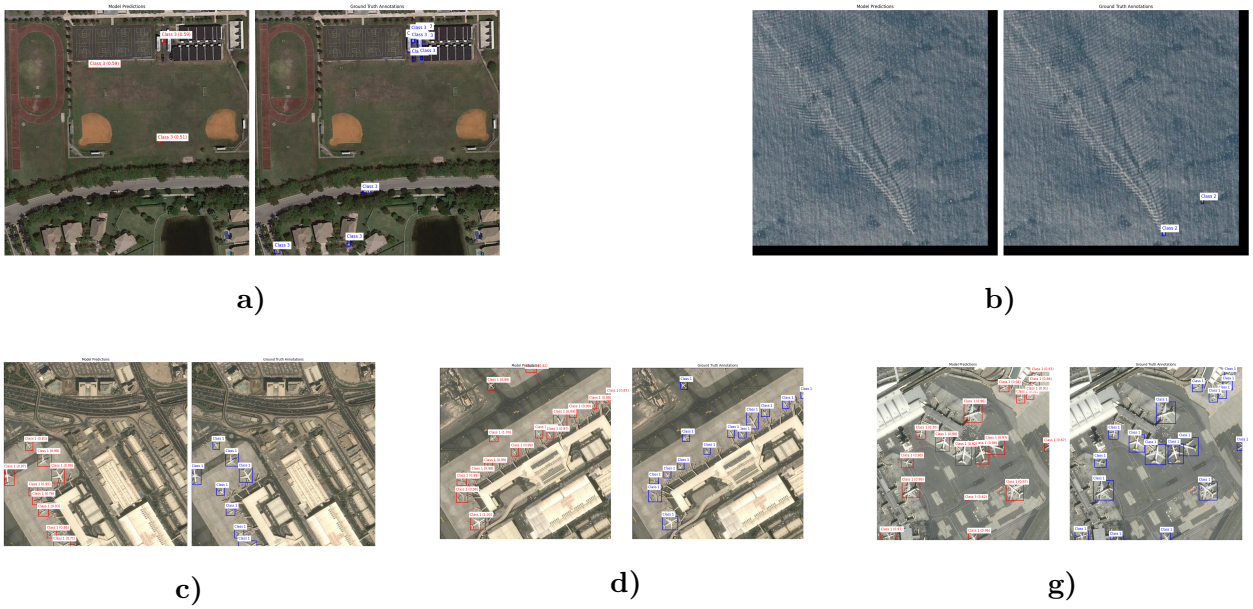


Figure 1: Fig. 1.1: FasterRCNN(Left Side of Each Image) vs Ground Truth Label(Right Side of Each Image). In d) as, we can notice, the ground truth for the top aircraft did not have a label but it is predicting it, for, b) The Faster R-CNN fails to detect the class for ships, for a) There are few missing predictions

While Faster R-CNN has demonstrated strong performance in object detection tasks, several challenges remain. In particular, we encounter issues such as:

(i) **Same Color/Camouflaged and Tiny Object Detection:** The model can struggle with detecting objects that share similar colors or patterns with their background, especially when the objects are small.

The custom SmallDetector was trained on,

a) R-CNN loss structure and,

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \sum_i \mathcal{L}_{\text{CE}}(p_i, p_i^*) + \lambda_{\text{reg}} \sum_i p_i^* \sum_j \mathcal{L}_{\text{smoothL1}}(t_{i,j} - t_{i,j}^*)$$

where:

$$\mathcal{L}_{\text{CE}}(p_i, p_i^*) = -\left[p_i^* \log p_i + (1 - p_i^*) \log(1 - p_i)\right]$$

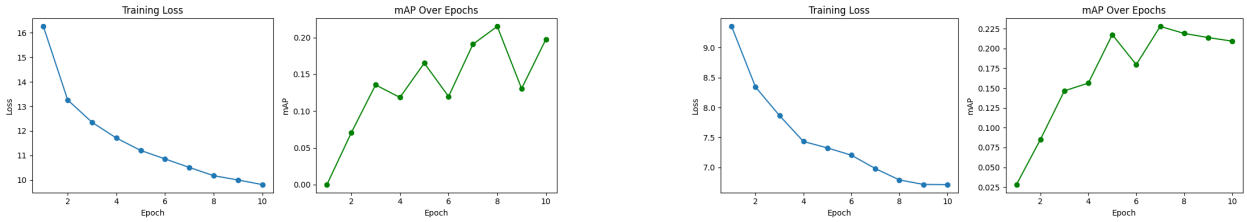
$$\mathcal{L}_{\text{smoothL1}}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}$$

with p_i as the predicted objectness score, p_i^* as the ground truth label, $t_{i,j}$ and $t_{i,j}^*$ as the predicted and ground truth bounding box parameters, and $\lambda_{\text{cls}}, \lambda_{\text{reg}}$ as weighting factors.

b) Normalized Wasserstein Distance (NWD) loss.

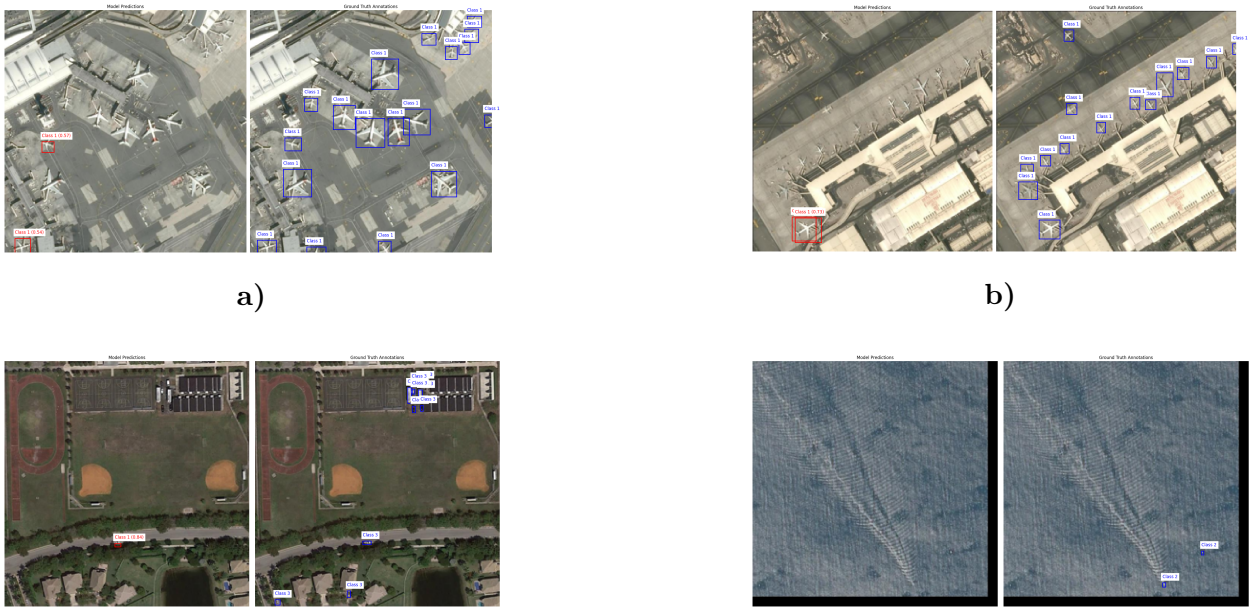
$$\mathcal{L}_{\text{total}} = \lambda_{\text{nwd}}\Big[(x_p - x_t)^2 + (y_p - y_t)^2 + (\sqrt{w_p} - \sqrt{w_t})^2 + (\sqrt{h_p} - \sqrt{h_t})^2\Big] + \lambda_{\text{conf}}(p_{\text{conf}} - 1)^2 + \lambda_{\text{noobj}}(p_{\text{conf}} - 0)^2 + \lambda_{\text{rka}}(A_p - A_t)^2$$

where $A_p = w_p h_p$, $A_t = w_t h_t$, and p_{conf} represents the confidence score.



(a) (b)

Figure 2: Loss and mAP plot for a) squared difference, b) Normalized Wasserstein Distance (NWD). The model was trained for 10 epochs with Adam optimizer with learning rate of 0.001



a) **b)**

c) **d)**

Figure 3: Fig. 1.2: SmallDetection(Left Side of Each Image) vs Ground Truth Label(Right Side of Each Image) using squared difference loss. In b) To resolve multiple detections of the same aircraft, a method is needed to select the most probable bounding box in case of overlap, ensuring accurate localization and minimal redundancy(techniques such as Non-Maximum Suppression (NMS) or Weighted Box Fusion (WBF)).

Our findings indicate that for detecting small changes, particularly in satellite imagery, the Normalized Wasserstein Distance (NWD) Loss and the Relative Kullback–Leibler Area (RKA) Loss outperform the traditional Faster R-CNN loss. The combined NWD + RKA loss exhibits greater

sensitivity to minor spatial and size variations, making it more effective in identifying tiny, camouflaged, or small-scale changes within an image. To further enhance the model's performance, we propose initializing it with additional parameters to improve its ability to capture subtle variations more effectively.

Model	**mAP**	**Loss**	**Model Size(in MB)**
Faster-RCNN	0.7314	0.40	153
SmallDetector(NWD)	0.225	6.84	0.250
SmallDetector(RCNN Loss)	0.20	9.91	0.250

Table 1: Comparison of models based on mAP(Mean Average Precision), Loss and Model Size

The SmallDetector is a compact model designed for object detection and localization. However, its limited capacity affects its ability to accurately detect objects in images, particularly in scenarios with high variability. This suggests that scaling up the parameter count of the SmallDetector (i.e., increasing its model size beyond 0.250 MB) could enhance its ability to capture subtle features, potentially improving the mAP and reducing the loss, though at the expense of computational efficiency.

3 Conclusion

The results indicate a substantial performance gap between Faster-RCNN and the SmallDetector variants. Faster-RCNN achieves a high mAP of 0.7314 with low loss, reflecting robust detection accuracy. Its large model size (153 MB) supports complex feature extraction and precise localization. In contrast, the SmallDetector models are extremely compact (0.250 MB) but exhibit significantly lower mAP (approximately 0.20–0.225). The elevated loss values for SmallDetector indicate a reduced capacity to capture subtle variations. This comparison suggests that model capacity, C, plays a critical role, as performance metrics improve with increasing C. Augmenting the parameter count in SmallDetector could potentially enhance its performance, i.e., improve mAP and reduce loss.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[2] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations (ICLR)*, 2015.