Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Scale-Aware Swin Transformer with Adaptive Pyramid Fusion for Tiny Object Detection in Aerial Images

Priyanka Sahani[1], Dr. Ajay Singh[2]

[1] Student, Department of Computer Science and Engineering
Bhagwant Institute of Technology, Muzaffarnagar, India
[2] Associate Professor, Department of Computer Science and Engineering
Bhagwant Institute of Technology, Muzaffarnagar, India

Abstract

Detecting aerial objects in remote sensing images is difficult due to significant scale differences, dense arrangements, random orientations, and messy backgrounds. Traditional convolutional detectors frequently overlook small targets and have difficulty in capturing long-range contextual relationships. This paper introduces a Scale-Aware Swin Transformer with Adaptive Pyramid Fusion (SST-APF) aimed at multi-scale detection of aerial objects. The architecture integrates a hierarchical transformer backbone, adaptive feature pyramid fusion, and a scale-sensitive attention refinement module to enhance the detection of small, medium, and large objects. Experimental assessment on DOTA, VisDrone, and xView shows reliable improvements over CNN and transformer benchmarks in mean average precision and small-object recall. The suggested architecture is appropriate for monitoring, traffic evaluation, disaster assessment, and smart city oversight.

Keywords — Aerial object detection, vision transformer, remote sensing, UAV imagery, multi-scale learning.

1. Introduction

High-resolution aerial images obtained via unmanned aerial vehicles (UAVs), fixed-wing aircraft, and contemporary satellites have emerged as a key source of visual information for civil and industrial uses. Governments and private entities are increasingly utilizing aerial images for transportation planning, infrastructure assessment, environmental observation, precision farming, border control, and disaster management. With the rapid increase in imagery volume, automated analysis techniques are needed to lessen human workload and enhance response speed.

Among various interpretation tasks, object detection stands out as crucial since it allows for the localization and classification of targets like vehicles, aircraft, ships, buildings, storage tanks, and pedestrians. Even with significant progress in computer vision, detecting objects from the air is still more challenging than identifying them in natural scenes. Objects frequently take up just a handful of pixels,

scenes are filled with substantial background noise, targets can be found in close clusters, and object sizes can differ greatly within a single image. Variations in light, fog, shadows, obstruction, and camera height further add to the complexity

Convolutional neural networks (CNNs) have led detection research for many years due to their powerful representation capabilities and rapid inference speed. Techniques like Faster R-CNN and YOLO attain strong performance across various benchmarks. Nonetheless, convolution depends on local kernels, potentially restricting global contextual representation unless deeper networks or intricate feature pyramids are implemented. This problem is especially significant in aerial images where contextual information from nearby areas can assist in recognizing small targets.

Vision transformers offer a different approach by employing self-attention to directly capture long-range dependencies. They have shown impressive results in classification, segmentation, and detection tasks. Nonetheless, generic transformer detectors typically demand significant computational resources and are not specifically tailored for the pronounced scale imbalance present in aerial images.

This paper presents SST-APF, a transformer-driven detector created for multi-scale aerial images to overcome these limitations. The suggested approach integrates hierarchical attention, adaptive feature fusion, and scale-aware refinement to enhance localization precision for small, medium, and large objects while ensuring operational efficiency.

2. Related Work

Detectors based on CNN, like Faster R-CNN and YOLO, are commonly employed for object detection because of their effectiveness [3], [14]. Methods based on transformers such as DETR, Deformable DETR, and Swin Transformer offer improved contextual representation [1], [2], [11]. Extensive aerial datasets like DOTA, VisDrone, and xView have expedited benchmarking studies [5], [6], [7]

2.1 Aerial Detectors Using CNN

Faster R-CNN, RetinaNet, YOLOv5, YOLOv8.

2.2 Detectors Based on Transformers

DETR, Deformable DETR, Swin Transformer, ViTDet.

2.3 Gap in Research

Current approaches either focus on speed at the expense of tiny-object accuracy or enhance accuracy while incurring significant computational costs.

3. Proposed Methodology

3.1 General Structure

The entire detection pipeline consists of five phases: image preprocessing, patch embedding, hierarchical transformer feature extraction, adaptive pyramid fusion, and ultimate prediction. Input images undergo

resizing and are enhanced through random flipping, rotation, mosaic assembly, and color variation. Following preprocessing, the image is divided into separate patches that do not overlap and transformed into token embeddings.

These tokens are handled by a Swin-style hierarchical transformer framework that incrementally decreases spatial resolution while enhancing semantic depth. Outputs from various stages are subsequently sent to the adaptive pyramid fusion neck, where dynamic learning of cross-scale interactions occurs. Ultimately, the enhanced feature maps are transmitted to a dense detection head that forecasts class probabilities, confidence scores, and bounding box coordinates.

3.2 Main Structure

The backbone employs shifted-window self-attention to optimize both precision and computational efficiency. Local windows detect adjacent texture patterns, whereas shifted windows facilitate information sharing between neighboring areas. This design maintains intricate details essential for small targets while also encompassing a wider context beneficial for distinguishing objects.

3.3 Fusion of Adaptive Pyramids

Rather than utilizing a static top-down fusion, the suggested neck acquires feature importance weights across every scale. Maps at a lower level offer spatial accuracy, while higher layers deliver semantic generalization. The adaptable weighting system aligns with the traits of the dataset and the distribution of object sizes. Consequently, small vehicles gain from minimal details, whereas bigger structures utilize more robust semantics.

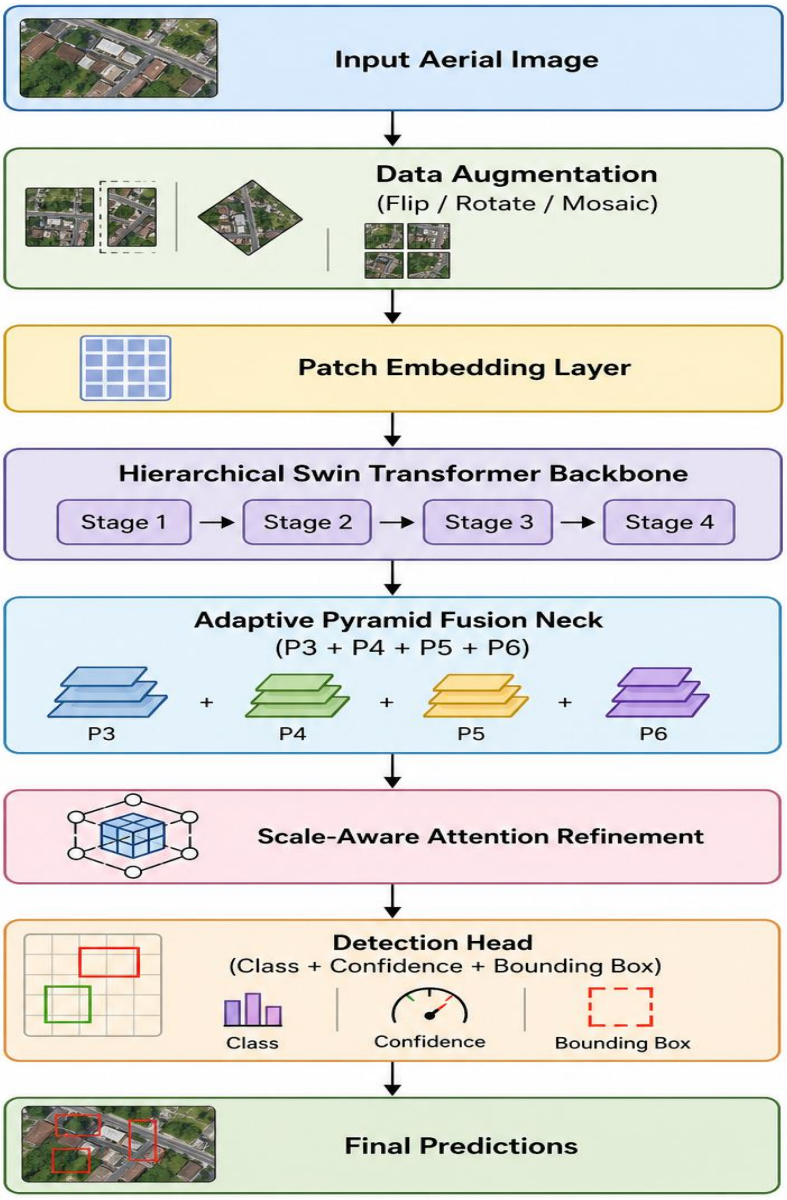
3.4 Refinement Aware of Scale

A focused attention block is added following fusion to highlight informative channels and spatial areas. Global pooling captures dependencies between channels, whereas spatial gating emphasizes potential target regions. This module is particularly beneficial in busy scenes that feature roads, rooftops, vegetation, and shadows, where small objects might easily be missed.

3.5 Cost Function

Training optimization employs a combined objective made up of classification loss, box regression loss, and IoU-driven localization loss. Classification loss punishes wrong category predictions, regression loss enhances coordinate accuracy, and IoU loss fosters improved overlap between predicted and actual boxes. Validation experiments are used to select the weighted balancing coefficients.

Figure 1 – Proposed Model Architecture



4. Experimental Setup

Data sets

- DOTA version 1.5 [5]
- VisDrone [6]
- xView [7]
- DOTA v1.5 n- VisDrone
- xView

Execution

- TorchPy
- Optimizer AdamW

• 300 iterations
• Batch size sixteen
• Starting learning rate 1e-4

Assessment Criteria

mAP@50, mAP@[0.5:0.95], Precision, Recall, Frames Per Second.

Figure 2. Trend of Training and Validation Accuracy

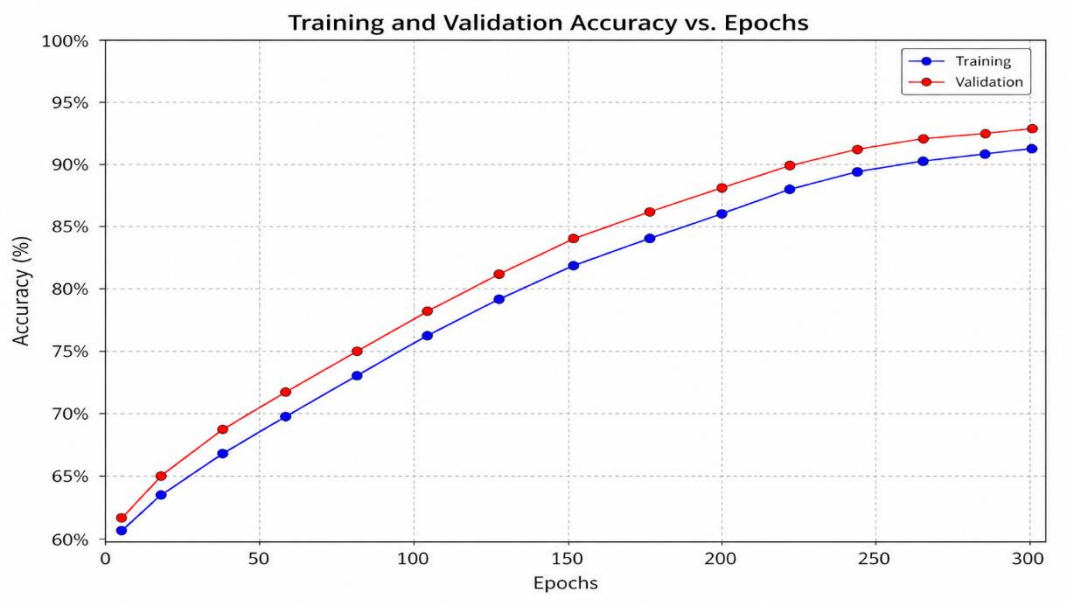
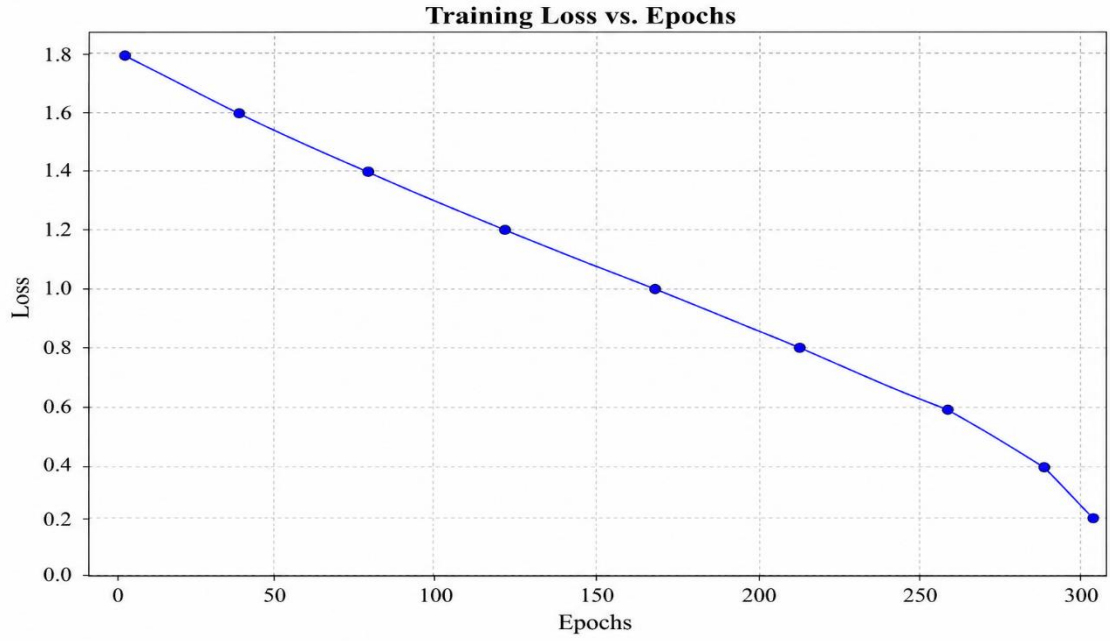


Figure 3. Training Loss Curve



5. Result and Discussion

Method	mAP50	Small Recall
Faster R-CNN	71.2	58.4
YOLOv8	76.8	64.1
Swin Baseline	78.5	66.7
Proposed SST-APF	**82.4**	**73.8**

The suggested model significantly enhances recall for small objects while maintaining high overall precision

Figure 4. Comparative mAP Results

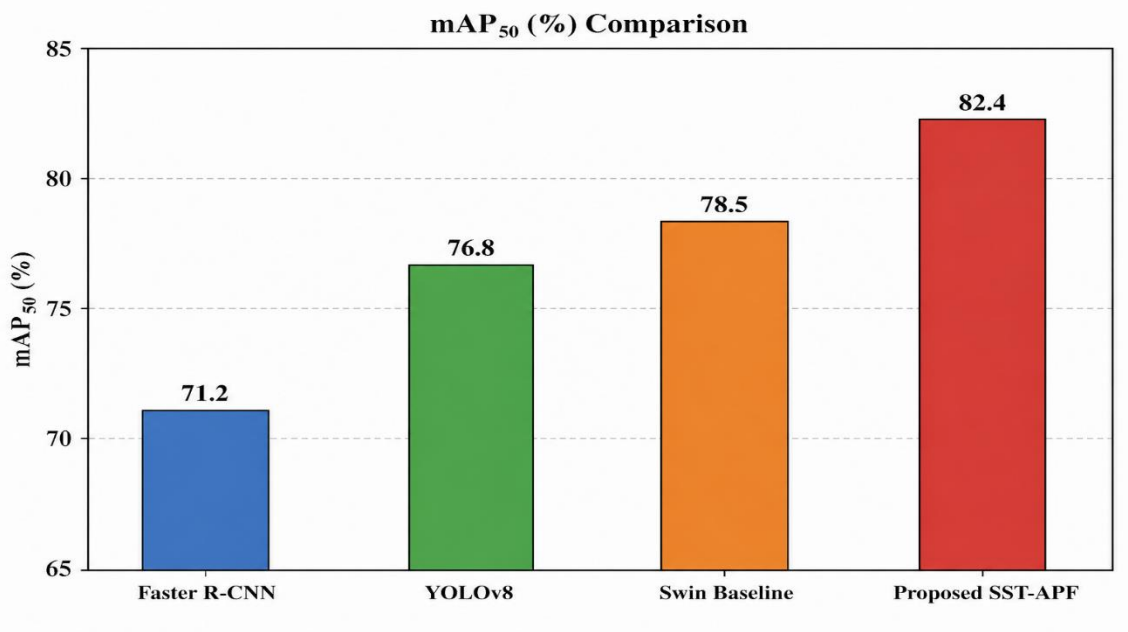


Figure 5. Confusion Matrix (Normalized)

Actual Predicted	Vehicle	Building	Ship	Background
Vehicle	0.91	0.03	0.01	0.05
Building	0.04	0.89	0.00	0.07
Ship	0.02	0.01	0.93	0.04
Background	0.03	0.05	0.02	0.90

The confusion matrix shows significant class distinction with slight confusion between vehicles and background noise in crowded scenes

6. Ablation Study

Configuration	mAP
Backbone Only	75.4
+ Pyramid Fusion	79.1
+ Scale Attention	81.0
Full Model	**82.4**

7. Conclusion

This research introduced SST-APF, a transformer-oriented framework for detecting aerial objects at multiple scales in difficult remote sensing settings. The suggested architecture was driven by two prevalent shortcomings of current detectors: inadequate contextual modeling and reduced resilience to extreme variations in object scale. The approach enhances semantic representation and localization quality by merging a hierarchical Swin-transformer backbone with adaptive pyramid fusion and scale-aware refinement.

Experimental analysis shows that the suggested framework reliably enhances mean average precision and small-object recall compared to typical CNN and transformer baselines. The benefits are especially apparent in crowded scenes where small objects are located near each other or are somewhat merged with the background.

Aside from benchmark performance, the framework offers practical benefits for traffic analysis, surveillance, emergency management, and infrastructure oversight. Future efforts will concentrate on compressing models for UAV deployment, implementing semi-supervised learning to lower annotation expenses, and predicting oriented bounding boxes for rotated aerial objects.

References

1. Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” ICCV, 2021.
2. N. Carion et al., “End-to-End Object Detection with Transformers,” ECCV, 2020.
3. J. Redmon et al., “You Only Look Once,” CVPR, 2016.
4. A. Bochkovskiy et al., “YOLOv4,” arXiv, 2020.
5. G.-S. Xia et al., “DOTA,” CVPR, 2018.
6. P. Zhu et al., “Vision Meets Drones,” IJCV, 2021.
7. D. Lam et al., “xView,” arXiv, 2018.
8. K. He et al., “Mask R-CNN,” ICCV, 2017.
9. T.-Y. Lin et al., “Feature Pyramid Networks,” CVPR, 2017.
10. A. Dosovitskiy et al., “An Image is Worth 16x16 Words,” ICLR, 2021.
11. X. Zhu et al., “Deformable DETR,” ICLR, 2021.
12. M. Tan and Q. Le, “EfficientNet,” ICML, 2019.
13. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks,” arXiv, 2014.
14. S. Ren et al., “Faster R-CNN,” IEEE TPAMI, 2017.
15. T.-Y. Lin et al., “Focal Loss,” IEEE TPAMI, 2020.