# TSFPN-YOLO: Object Detection under Low-Altitude Traffic Surveillance Using a Twin-Stream Feature Pyramid Network

**Shengjie Feng[a], Kaixuan Cui[b], Shu Guo[c], Bingcheng Jiang[d,*]**

*Guilin University of Electronic Technology, Guilin, China*
*[a]ikunkukuku@gmail.com, [b]1316698892@qq.com, [c]2602340770@qq.com,*
*[d]jiangbingcheng@guet.edu.cn*
*[*]Corresponding author*

**Abstract:** *With the swift advancement of unmanned aerial vehicle (UAV) technology, low-altitude traffic surveillance has become crucial for intelligent traffic management and is extensively used in military reconnaissance, police suspect tracking, and large-scale target searches. Twin-Stream feature fusion methods can enhance detection accuracy and system robustness. However, current methods face challenges in detection accuracy and real-time performance in dynamic, low-altitude scenarios. This paper introduces a novel object detection approach using a modified TSFPN_Concat structure to optimize YOLOv8 for these tasks. We present an improved Twin-Stream Feature Pyramid Network fusion mechanism that boosts detection accuracy and adaptability in complex scenarios. To validate the effectiveness of the proposed model, we refined and curated a portion of the dataset [1] published in 2020. Compared with mainstream methods such as YOLOv5, YOLOX, Faster R-CNN, and Mask R-CNN, our experimental results demonstrate that the improved TSFPN-YOLO outperforms the original YOLOv8[2] in multiple metrics. Notably, it achieves a 45.2% mAP@50%, surpassing other comparison models and demonstrating remarkable performance improvement. Research indicates that the improved TSFPN_Concat structure can effectively address the insufficient multi-scale feature fusion problem in low-altitude object detection, thereby enhancing the accuracy in low-altitude traffic surveillance tasks.*

**Keywords:** *Object Detection, Twin-Stream Feature Pyramid Network (TSFPN), Low-Altitude Traffic Surveillance, YOLOv8*

## 1. Introduction

Currently, traditional ground-based object detection remains the mainstream research focus in the field of object detection. However, with the continuous advancement of deep learning, the rapid development of unmanned aerial vehicle (UAV) technology, the increase in building heights, and the growing sophistication of road monitoring infrastructure, aerial-view object detection has begun to attract increasing attention from researchers. In particular, low-altitude traffic surveillance has emerged as a key component in modern intelligent transportation systems, offering broad prospects for practical applications. Low-altitude monitoring not only facilitates the intelligent management of urban traffic but also demonstrates significant potential in fields such as military reconnaissance, police pursuit of suspects, and large-scale target searches. However, due to the complexity and dynamic nature of low-altitude environments, object detection tasks in such scenarios are especially challenging. Traditional ground-based object detection methods often fail to meet the practical requirements of aerial environments. Moreover, low-altitude scenarios are fraught with multiple issues. The small sizes of targets, scarcity of feature information, and intense background interference all conspire to render object detection an even more arduous task. Consequently, augmenting the detection capabilities for diminutive targets and complex, dynamic scenes within low-altitude environments has emerged as a crucial research conundrum that demands immediate attention and innovative solutions.

The rise of deep learning has led to revolutionary progress in object detection tasks. Methods based on convolutional neural networks (CNNs) have become the mainstream approach. From the original R-CNN series [3] to the YOLO (You Only Look Once) series, deep learning techniques have achieved remarkable breakthroughs in both detection accuracy and real-time performance. The YOLO series, particularly YOLOv8, stands out as a state-of-the-art model in object detection, offering high speed and accuracy and being widely applied across various computer vision tasks. Nevertheless, while YOLOv8 performs well under traditional conditions, it still faces challenges in low-altitude environments characterized by complex backgrounds, rapidly moving targets, and multi-scale detection requirements.

In low-altitude traffic surveillance, the accuracy and real-time performance of object detection algorithms are critical factors in determining system effectiveness. Although existing methods—such as the YOLO series, Faster R-CNN, and Mask R-CNN—perform admirably on standard datasets, they still face multiple challenges when confronted with highly dynamic scenes and complex backgrounds, especially in UAV imagery captured at low altitudes. These challenges include diminished detection accuracy for fast-moving targets, interference from complicated backgrounds, and difficulties in locating multi-scale objects.

In light of these issues, this study proposes an object detection approach based on an improved Twin-Stream Feature Pyramid Network (TSFPN) structure. We optimize the YOLOv8 model by employing an enhanced TSFPN fusion mechanism, utilizing multi-level feature fusion to further improve YOLOv8's object detection performance in low-altitude traffic surveillance. More specifically, the improved TSFPN structure effectively integrates feature information from different scales and layers, thereby enhancing detection accuracy and robustness in low-altitude environments. Through experimental validation on the "A Multi-modal Unmanned Aerial Vehicle Dataset for Low-Altitude Traffic Surveillance" published on arXiv in 2020, we tested our improved YOLOv8 model. The results show that the improved model outperforms traditional detection methods such as YOLOv5, YOLOX, Faster R-CNN, and Mask R-CNN across multiple evaluation metrics. Notably, it achieves a significant improvement with a 45.2% mAP@50%.

The main contributions of this paper are as follows:

1) We propose an improved TSFPN fusion structure that utilizes Twin-Stream weighted fusion of multi-level features, significantly enhancing feature retention efficiency during concatenation.

2) By incorporating multi-scale feature fusion techniques, we substantially improve the adaptability of the YOLOv8 model to complex scenarios.

3) Through experiments conducted on a refined version of the "A Multi-modal Unmanned Aerial Vehicle Dataset," we validate the effectiveness of the improved model, which outperforms mainstream detection methods on multiple performance indicators.

## 2. Related Works

### *2.1 Object Detection*

Object detection is a core task in computer vision, aiming to identify and locate objects within images. Traditional object detection methods, such as those based on sliding windows, typically rely on handcrafted features (e.g., edges, textures, colors). However, these methods often perform poorly when confronted with complex backgrounds and multi-scale objects. With the rise of deep learning, convolutional neural network (CNN)-based approaches have gradually become the mainstream. By autonomously extracting features, these methods avoid the complexity of feature engineering while exhibiting greater flexibility in handling images with complex backgrounds.

Typical deep learning-based object detection methods can be divided into two categories: two-stage detectors (e.g., Faster R-CNN) and one-stage detectors (e.g., YOLO, SSD). Two-stage detectors first generate region proposals (via a Region Proposal Network, RPN) and then perform classification and regression. While they achieve high detection accuracy, their computational cost is substantial, making them suitable for scenarios that require high precision but have lower real-time constraints. In contrast, one-stage detectors directly generate predictions from the image, resulting in faster detection speeds and making them more suitable for real-time applications. The YOLO series, in particular, represents the mainstream in one-stage detection. In recent years, research in object detection has increasingly focused on handling more complex scenes, a greater variety of objects, and higher real-time requirements. Against this backdrop, multi-scale feature fusion techniques and enhanced feature representations have become important research directions.

For low-altitude object detection tasks, Dong et al. [4] utilized a triple attention mechanism and a lightweight feature fusion network, along with an innovative downsampling module, to propose an improved YOLOv8 algorithm under a UAV low-altitude perspective. Xu et al. [5] introduced an improved YOLOv5s-based model as a lightweight solution to low-altitude object detection problems on the battlefield while maintaining detection accuracy. Qing-bang Shi et al. [6] constructed a UAV flight posture dataset through self-captured images and online sources and introduced the YOLOv4 algorithm, marking the first use of YOLOv4 in low-altitude UAV object detection.

### *2.2 YOLOv8*

YOLO (You Only Look Once) is one of the most influential algorithms in object detection. By transforming object detection into a regression problem, YOLO allows detection to be completed in a single forward pass, significantly improving detection speed. Since its initial release in 2015, the YOLO series has undergone continuous updates and iterations, gradually becoming a benchmark in object detection due to its excellent detection accuracy and high efficiency.

YOLOv8 stands out as an excellent version of the YOLO series, retaining the overall framework while optimizing feature extraction and fusion mechanisms relative to earlier YOLO versions. It employs a more efficient backbone network and more complex feature fusion strategies, demonstrating greater robustness in handling both large-scale and small-scale objects. Additionally, YOLOv8 incorporates new loss functions and regularization techniques to improve performance in multi-task learning. It also leverages Feature Pyramid Networks (FPN), Path Aggregation Networks (PAN), and new annotation tools, hinting at improvements in the labeling process.

Although YOLOv8 achieves good performance on most standard datasets, it still faces challenges in low-altitude environments, especially in complex and dynamic scenes. In this study, we propose an improvement to YOLOv8 by introducing an enhanced TSFPN structure. Through more efficient multi-scale feature fusion, we enhance object detection performance in low-altitude traffic surveillance scenarios.

### *2.3 Feature Fusion*

Feature fusion techniques play a critical role in object detection. Especially in complex scenes, relying on a single feature level often fails to capture all the necessary information. The goal of feature fusion is to combine feature maps from different layers so that high-level semantic features and low-level detail features can complement each other, thereby improving object detection performance

Feature Pyramid Networks (FPN) represent one of the earliest multi-level feature fusion methods. Through top-down feature propagation, FPN passes high-level semantic information to lower layers to improve the detection accuracy of small-scale objects. Its working principle reduces the spatial resolution of input images while increasing the number of feature channels. Although FPN demonstrates strong capability in multi-scale object detection, it is inherently limited by its unidirectional information flow, resulting in some constraints on feature fusion weighting.

To address the limitations of FPN, PANet (Pyramid Attention Network) further refines the feature fusion structure by enhancing information flow between high-level and low-level features. By introducing a bottom-up path aggregation network, PANet captures more spatial context information, improving model detection accuracy and adaptability.

In this study, we propose an improved TSFPN (Twin-Stream Feature Pyramid Network) structure that employs cross-scale connections and weighted feature fusion. This approach enables more efficient learning of shared features across different layers, thereby improving the YOLOv8 model’s performance in low-altitude traffic surveillance. Particularly in complex, dynamic environments, our method significantly enhances detection accuracy and real-time capabilities.

## 3. Proposed Method

In this section, we first introduce the overall framework of the TSFPN-YOLO architecture, then describe the proposed TSFPN feature fusion module in detail, and finally analyze its advantages.

### *3.1 Method Overview*

While YOLOv8 achieves satisfactory detection accuracy in traditional ground-based static viewpoints, its performance in low-altitude surveillance scenarios, where objects are rapidly changing, is somewhat lacking. One primary cause of this shortfall is information loss during multi-level feature fusion. To address this issue, we propose a new network architecture named TSFPN-YOLO, which integrates a multi-scale feature fusion module (TSFPN) into the YOLOv8 detection framework, as shown in Fig.1.

In the original YOLOv8 architecture, feature fusion across layers is carried out by channel-wise concatenation with equal weighting. This approach can cause the model to fail to fully leverage the features from layers that contain richer information, while still incorporating less informative layers. As a result, information is lost and useless learning occurs. TSFPN-YOLO employs an effective Twin-Stream, cross-scale connection and weighted feature fusion strategy to enhance feature learning. This, in turn, significantly improves the accuracy of YOLOv8 for target detection in low-altitude surveillance scenarios

### *3.2 Network Architecture*

The main idea behind TSFPN is the effective use of bidirectional cross-scale connections and weighted feature fusion. In TSFPN, each node receives inputs from at least two other nodes. We remove nodes that have only a single input edge because a node with a single input does not provide true multi-level feature fusion, thus offering limited contribution to the feature network structure. This is why our network is called “Twin-Stream” (TS), it

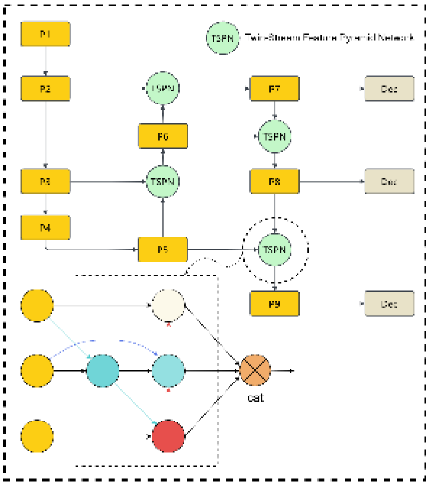

*Figure 1: The architecture of TSFPN-YOLO, explains the TSFPN feature fusion module to enhance the feature extraction capability of the model.*

Consistently utilizes dual-stream (or multiple input) paths. Additionally, we add an extra edge from the original input node to the output node, thereby achieving richer feature integration without increasing the number of feature layers used.

Prior methods treat all input features equally, assuming identical contribution to the final output. However, our research reveals that different channels and different resolutions contribute unequally to the fused features. To address this, the TSFPN module assigns an additional weight to each input feature and learns its importance via backpropagation. To prevent instability and ensure that feature representations remain within a reasonable range, we employ weight normalization, ensuring that all feature weights sum to 1. Specifically, we apply the softmax function to each weight and add a small bias term to avoid numerical instability such as division by zero.

$$O = \Sigma_i \frac{w_i}{\epsilon + \Sigma_j w_j} \tag{1}$$

### *3.3 Advantages of the Proposed Method*

Traditional FPN is constrained by unidirectional information flow and cannot effectively integrate features across different layers. Although PANet improves upon FPN by adding a bottom-up path aggregation network, it still struggles to fully extract useful features among different layers.

In contrast, TSFPN adopts a bidirectional cross-scale feature fusion approach, ensuring that each node receives information from more than two layers. This enables the network to thoroughly learn multi-level features. Meanwhile, the use of weighted feature layers allows the network to learn which features are more useful. By employing normalization strategies for the weights, TSFPN ensures stable and accurate feature fusion, avoiding the instability that can arise during training.

Overall, incorporating TSFPN endows the network model with greater robustness. It maintains high detection accuracy even in uncertain and complex environments. In the low-altitude UAV traffic perspective used in this study, facing highly dynamic targets and complex backgrounds, TSFPN-equipped YOLOv8 still performs excellently, demonstrating significant improvements in both accuracy and real-time performance.

## 4. Experiments

### *4.1 Experimental Data and Implementation Details*

**Dataset**: In this study, we use the AU-AIR dataset to verify the effectiveness of the TSFPN model. The AU-AIR dataset is specifically designed for object detection and image processing research in low-altitude traffic scenarios. It contains 32,823 images annotated with 8 object categories and includes varying flight altitudes between 5 and 30 meters as well as camera angles between 45° and 90°. To accelerate training and shorten the verification time of the proposed model, we refine the AU-AIR dataset by randomly selecting 50% of the data for training and removing three categories with fewer annotations. Of this refined AU-AIR dataset, 80% is used as the training set, and 20% is used as the test set.

**Implementation Details**: All training and testing are conducted with input images resized to 640×640. We set the batch size to 16 and use Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01. Each model is trained for 50epochs on an NVIDIA TITAN Xp GPU using the PyTorch framework.

**Evaluation Metrics**: We evaluate the effectiveness of different models at IoU=0.5 using the mAP (mean Average Precision) metric. mAP represents the average detection precision across all object categories. A higher mAP value indicates stronger detection capabilities. The calculation formula is as follows:

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{2}$$

Where C is the number of categories, and AP denotes the Average Precision for each category, calculated as the area under the Precision-Recall (P-R) curve.

### *4.2 Comparison Experiments*

In this section, we compare our model with various object detection models on the AU-AIR dataset. Table 1 shows the experimental results. Compared with two-stage object detection models, TSFPN-YOLO improves mAP@0.5 by 5.3% over Mask R-CNN and by 6.2% over Faster R-CNN. Compared with one-stage object detection models, our model also demonstrates improvements, surpassing YOLOv8 by 0.4%, YOLOv5 by 0.6%, and YOLOX by 0.8%.

*Table 1: Comparison of MAP results of TSFFPN-YOLO and other current mainstream object detection models on AU-AIR dataset*

| Model | Backbone | Size | Map@0.5/% |
|---|---|---|---|
| Yolov5 | CSPDarknet | 640×640 | 44.6 |
| Yolox | CSPDarknet | 640×640 | 44.4 |
| Mask-RCNN | ResNet50 + FPN | 640×640 | 39.9 |
| Faster-RCNN | ResNet50 + FPN | 640×640 | 39.0 |
| Yolov8 | DarkNet-53 | 640×640 | 44.8 |
| **TSFPN-Yolo** | **DarkNet-53** | **640×640** | **45.2** |

### *4.3 Visualization on Low-Altitude Datasets*

Fig.2. compares the detection results of TSFPN-YOLO with other mainstream methods, providing a direct visual comparison. The proposed algorithm accurately identifies targets under different altitudes and viewing angles, whereas other algorithms may exhibit errors, missed detections, or incorrect identifications.

In summary, our algorithm achieves high detection accuracy and strong model robustness for low-altitude object detection. Even in highly dynamic UAV perspectives, it maintains excellent recognition performance.

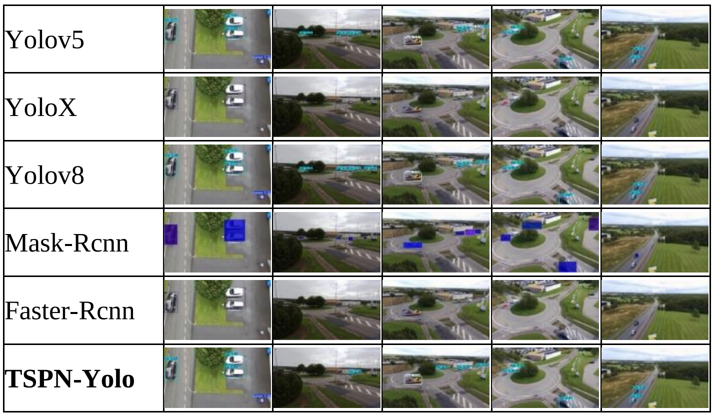

*Figure 2: The visual comparison between TSFPN and other mainstream target detection methods proves the target detection performance of TSFPN-YOLO under low altitude dynamic conditions.*

## 5. Conclusion

This paper proposes TSFPN-YOLO, a method aimed at improving the accuracy of object detection from low-altitude surveillance perspectives, providing a viable optimization solution for low-altitude object detection tasks. By refining the feature fusion module of YOLOv8 through cross-scale connections and weighted feature fusion, the model's ability to recognize features during layer fusion is significantly enhanced. Experimental results demonstrate that TSFPN-YOLO substantially improves object detection accuracy from a low-altitude viewpoint. Compared to various mainstream CNN-based detectors, it achieves considerable progress, attaining a 45.2% mean Average Precision (mAP) on the AU-AIR dataset. The findings not only foster innovative thinking within the field of computer vision but also mark a new breakthrough in low-altitude traffic surveillance. Moreover, this research can offer reliable reference value for urban environmental management, aerial cinematography, military search operations, and related domains.

## References

*[1] Bozcan I, Kayacan E. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020: 8504-8510.*

*[2] Sohan M, Sai Ram T, Reddy R, et al. A review on yolov8 and its advancements[C]//International Conference on Data Intelligence and Cognitive Informatics. Springer, Singapore, 2024: 529-545.*

*[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.*

*[4] Dong, Y.-B., Zeng, H., & Hou, S.-J. LMUAV-YOLOv8: A Lightweight Network for Low-Altitude UAV Visual Object Detection [J/OL]. Computer Engineering and Applications, 1–21 [2024-12-11]. Retrieved from http://kns.cnki.net/kcms/detail/11.2127.tp.20241030.1319.008.html.*

*[5] Xu, T.-N., Gao, A., Chen, Y.-C., et al. A Lightweight Recognition Method for Low-Altitude Battlefield Targets [J/OL]. Journal of Ordnance Engineering, 1–12 [2024-12-11]. Retrieved from http://kns.cnki.net/kcms/detail/11.2176.TJ.20240816.1339.002.html.*

*[6] Shi Q, Li J. Objects detection of UAV for anti-UAV based on YOLOv4[C]//2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT. IEEE, 2020: 1048-1052.*