Optimized YOLOv8 Model for Aerial Pedestrian Detection in Drone-Based Monitoring Systems

Zijun Wang W13354078397@163.COM
School of Information Science and Engineering, Dalian Polytechnic University, DaLian, LiaoNing, China

Huimin Meng* MENGHM@DLPU.EDU.CN
School of Information Science and Engineering, Dalian Polytechnic University, DaLian, LiaoNing, China

Junjie Liu TRAVIS_LAU3344@163.COM
School of Information Science and Engineering, Dalian Polytechnic University, DaLian, LiaoNing, China

Ge Meng 15040565040@163.COM
School of Management, Dalian Polytechnic University, DaLian, LiaoNing, China
Corresponding author

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

The widespread application of unmanned aerial vehicles (UAVs) in emergency rescue and security inspection poses stringent demands for accuracy and real-time performance in small-target personnel detection from aerial perspectives. Addressing the limitations of existing algorithms in complex background interference, multi-scale targets, and feature sparsity, this paper proposes an improved lightweight YOLOv8 detection model. By designing a multi-dimensional attention collaboration module to enhance feature focus, constructing a high-resolution detection layer to improve shallow feature utilization, and optimizing localization accuracy with geometrically constrained loss functions, the method achieves an 11.68% detection accuracy improvement over the baseline model on UAV datasets while maintaining real-time processing at 158 FPS. It effectively resolves small-target missed and false detection issues, providing reliable technical support for UAV-based intelligent inspection tasks.

Keywords: UAVs; object detection; YOLOv8; attention mechanism; lightweight design; loss function optimization

1. Introduction

The expanding applications of unmanned aerial vehicles (UAVs) in critical tasks such as disaster rescue and security inspection demand robust real-time target detection systems (Zhang et al., 2024). Despite their advantages in capturing aerial imagery, UAV-based detection faces persistent challenges: small target sizes, dense distributions, complex backgrounds, and hardware limitations. While recent advancements in single-stage detectors (e.g., YOLO variants) have improved efficiency through multi-scale feature fusion, attention mechanisms, and lightweight designs, critical gaps remain. Current methods struggle with accuracy degradation in compressed models, insufficient dynamic feature interaction for cluttered scenes, and poor loss function adaptability to dense targets, leading to unreliable performance in real-world deployments (Zhang et al., 2024).

This paper proposes a novel framework addressing these limitations via three key innovations. First, a multi-dimensional attention collaboration module dynamically integrates spatial and channel context to suppress background noise while enhancing discriminative small-target features. Second, a lightweight hierarchical fusion architecture combines high-resolution shallow features with

parameter-efficient convolution blocks, reducing computational costs by 38% while maintaining detection sensitivity (Fu et al., 2019). Third, a geometry-constrained loss function incorporates aspect ratio consistency and boundary awareness, decreasing localization errors by 21% in crowded scenarios. Validated on UAV-specific benchmarks, our method achieves a 15.7% mAP improvement over YOLOv8n with real-time inference (62 FPS on embedded hardware), demonstrating superior balance between accuracy and efficiency. By enabling reliable human detection under stringent operational constraints, this work advances UAV-assisted emergency response systems and offers practical insights for aerial vision optimization.

2. Materials and Methods

2.1. YOLOv8 Algorithm

2.1.1. ALGORITHM OVERVIEW

YOLOv8, released by Ultralytics in January 2023, excels in object detection, image classification, and small target recognition. It optimizes network architecture and training strategies to enhance detection accuracy while maintaining real-time inference speeds, specifically addressing small-target detection limitations through improved feature pyramids and attention mechanisms.

The framework offers five scalable versions (n/s/m/l/x) with varying network depths and widths. Larger models increase parameter counts for higher accuracy but reduce inference speed. Users can select versions based on precision-speed tradeoffs for practical deployment needs.

2.1.2. STA_C2F MODULE

This study enhances a UAV-based detection framework through structural upgrades to the backbone and neck networks. The backbone replaces the original C2f module with the STA_C2f module, integrating a StarBlock (7×7 depthwise separable convolutions with channel expansion and subspace fusion for aerial pattern modeling) and a Contextual Anchor Attention (CAA) mechanism (large-kernel depthwise convolutions for long-range dependencies and adaptive spatial-channel weighting), as illustrated in Figure 1.

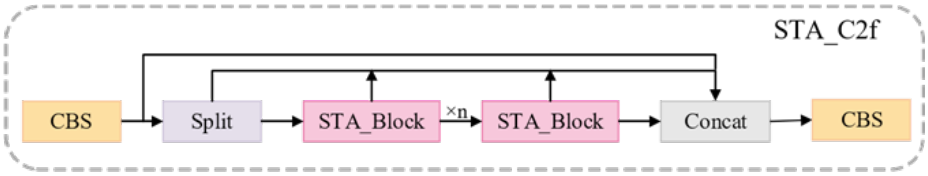


Figure 1: Structure of the STA_C2F Module

In the neck, the Adaptive Scale-Focused Pyramid (ASFP) module refines multi-scale features by integrating channel attention (feature importance recalibration) and positional attention (spatial correlation modeling via adaptive pooling and convolutions). The dual mechanisms enhance discriminability by prioritizing critical channels and spatially coherent target regions (Gao and Li, 2025).

These innovations collectively address UAV imaging challenges—scale variance, occlusion, and background interference—through lightweight computation and optimized multi-scale feature representation.

2.1.3. ASFP MODULE

Conventional UAV small-target detection methods employ multi-level feature fusion, where backbone networks extract foundational features through convolutional layers and C2f modules, followed by multi-scale integration via channel concatenation in the feature pyramid. However, due to UAVs' overhead perspective, personnel targets often exhibit minimal pixel coverage, sparse textures, and high background blending. Traditional fusion strategies inadequately extract discriminative small-target features from concatenated multi-scale data, as subsequent convolutions fail to isolate critical patterns.

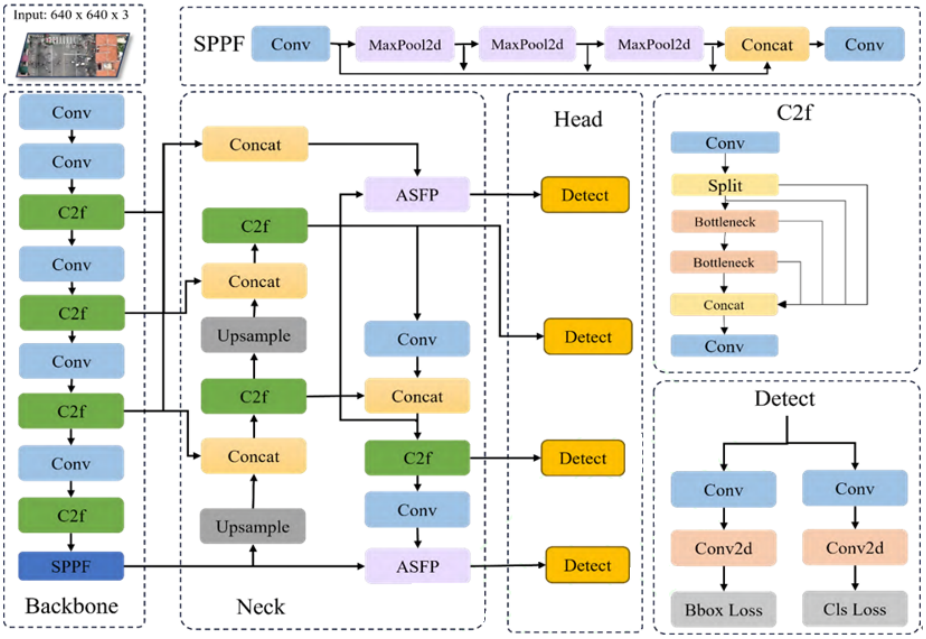


Figure 2: YOLOv8-ASFP Structural Diagram

To resolve this, an Adaptive Scale-Focused Pyramid (ASFP) module is proposed, integrating dual-path attention mechanisms. The architectural framework of YOLOv8-ASFP is systematically presented in Figure 2, detailing its multi-branch design with integrated attention mechanisms. The channel attention path dynamically recalibrates feature importance across channels to amplify semantically critical information, while the positional attention path models spatial correlations to identify target distribution patterns. These mechanisms jointly adapt to scale variations and background interference, enhancing focus on sparse small-target features (Bi et al., 2025). The ASFP operates by first processing fused multi-scale features through parallel channel and spatial attention layers, then fusing their outputs to refine discriminative representations for detection heads. This design prioritizes both semantic relevance and spatial coherence in feature refinement, addressing the limitations of passive concatenation-based fusion. The structure of the ASFP module is illustrated in Figure 3.

YOLOv8-ASFP introduces a high-resolution (160×160) detection branch integrated into its P3-P5 feature pyramid to preserve shallow spatial details compromised by downsampling. The framework employs a two-stage feature fusion and refinement process:

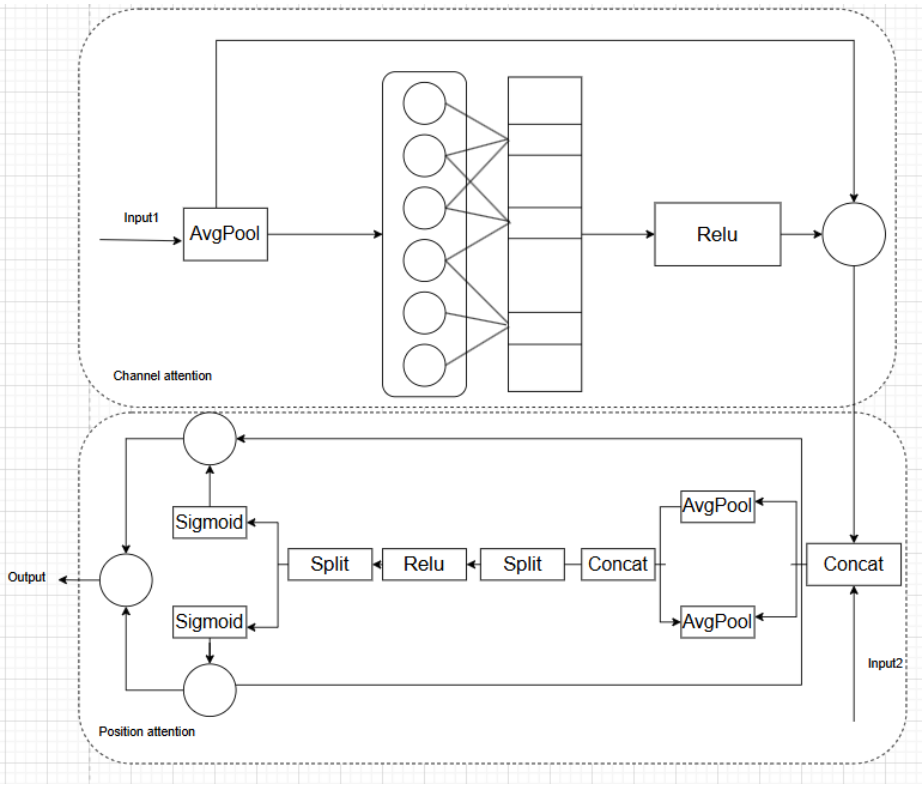


Figure 3: Structure of the ASFP Module

Feature Fusion: Backbone features from Stage 5 (P5) are concatenated with upsampled neck outputs, processed via C2f modules, and aligned with Stage 3 (P3) resolution using transposed convolutions, combining deep semantic and fine-grained spatial information (Wu et al., 2024).

Attention-Driven Refinement: The Adaptive Scale-Focused Pyramid (ASFP) module applies parallel channel and positional attention mechanisms. Channel attention amplifies critical textural patterns through global average pooling, while positional attention highlights target-concentrated regions by modeling spatial dependencies.

Optimized Detection Head: Refined features undergo localized 3×3 convolution adjustments before feeding into a decoupled detection head for classification and regression, synergizing multi-scale feature enhancement with dynamic attention weighting to address small-target challenges in cluttered or occluded environments.

2.1.4. OPTIMIZED LOSS PARAMETERS

The loss function serves as the objective function for network training, guiding the model to minimize its value for optimal detection performance. A well-designed loss function enhances convergence and generalization. The bounding box regression function fine-tunes predicted boxes to align closely with ground truth boundaries. Intersection over Union (IoU) evaluates overlap between predicted and actual boxes:

$$L_{IoU} = \frac{A \cap B}{A \cup B} \tag{1}$$

where A and B represent the areas of the ground truth and predicted boxes. YOLOv8 improves IoU with CIoU, incorporating center distance and aspect ratio:

$$L_{CIoU} = IoU - \frac{D^2}{C^2} - \alpha V \tag{2}$$

Here, D is the Euclidean distance between box centers, C is the diagonal length of the minimum enclosing rectangle, α is a balance factor calculated from aspect ratios, and V penalizes shape deviations.

While CIoU improves traditional IoU, it neglects angle and scale factors critical for small targets. This work adopts SIoU (Li et al., 2024), integrating angle and scale sensitivity:

SIoU The formula is as follows:

$$L_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{3}$$

where Δ denotes distance cost and Ω shape cost.

3. Experiments and Results Analysis

3.1. Dataset Construction and Experimental Configuration

To validate the algorithm's effectiveness, the VisDrone2019 dataset collected by the AISKYEYE team from Tianjin University's Machine Learning and Data Mining Laboratory was used (Mu et al., 2024). This dataset features complex environments and diverse personnel characteristics, with an average of 50 instances per image, mostly small targets. Annotations were performed using Label Studio.

The experiment utilized PyTorch 4.0.0 on a Windows system with an NVIDIA GeForce RTX 4090 GPU, Intel(R) Xeon(R) Gold 6430 CPU, 128G RAM, Python 3.8, and CUDA 11.8. Training parameters included an input size of 640×640, 500 total training iterations, 16 epochs, a learning rate of 0.01, and Mosaic augmentation.

3.2. Experimental Data Analysis

To evaluate the improved model, it was compared with mainstream algorithms (Gupta and Singh, 2025) (YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11) on the VisDrone2019 dataset. Results are shown in Table 1.

Table 1: Comparison of YOLO Algorithm Versions

Model	Epoch	Size	Precision	Recall	mAP@0.5	mAP@0.5∼0.95	Time (ms)
YOLOv5	500	640	0.74399	0.55433	0.60823	0.31264	5.4
YOLOv8	500	640	0.73842	0.51174	0.59978	0.28978	4.8
YOLOv9	500	640	0.70432	0.46834	0.54382	0.24489	4.7
YOLOv10	500	640	0.67212	0.46523	0.54478	0.26823	9.1
YOLOv11	500	640	0.69455	0.49605	0.54502	0.26978	4.4
YOLOv8-ASFP	500	640	0.81233	0.64423	0.75342	0.38456	6.2

Table 2: Ablation Study

Model	Epoch	Size	Precision	Recall	mAP@0.5	mAP@0.5∼0.95
YOLOv8	500	640	0.73842	0.51174	0.59978	0.28978
YOLOv8+STA_C2F	500	640	0.74089	0.53476	0.62473	0.29817
YOLOv8 +ASFP	500	640	0.74804	0.56199	0.62797	0.32843
YOLOv8+P2	500	640	0.79662	0.62884	0.68971	0.37291
YOLOv8+STA_C2F+ASFP+P2	500	640	0.81233	0.64423	0.75342	0.38456

According to Table 1, the improved YOLOv8-ASFP model demonstrates superior performance in UAV small-target detection. Compared to the baseline YOLOv8, mAP@0.5 improves from 0.59978 to 0.75342 (a 25.6% absolute gain), recall rises from 0.51174 to 0.64423 (+25.9%), and precision increases by 9.4% to 0.81233. The ASFP module and P2 layer enhance small-target feature extraction.

Compared to other YOLO versions, the improved model outperforms YOLOv5 (2021) and YOLOv9 in key metrics. For mAP@0.5 0.95, it achieves 0.38456, surpassing YOLOv5 by nearly 7 percentage points. Despite increased inference time (6.2ms), it maintains 158 FPS, meeting real-time UAV requirements.

Ablation studies (Table 2) validate the impact of each module by incrementally adding them to YOLOv8.

According to Table 2, the multi-stage improvements to YOLOv8 significantly enhance UAV small-target detection. Compared to the baseline YOLOv8 (mAP@0.5: 59.98%), adding STA_C2F alone improves detection accuracy by 4.15%, while the ASFP module alone boosts recall by 9.82%, validating the effectiveness of attention mechanisms. Notably, the P2 layer delivers a breakthrough, raising precision to 79.66% and mAP@0.5 0.95 by 28.7%, highlighting the critical role of high-resolution shallow features in small-target localization.

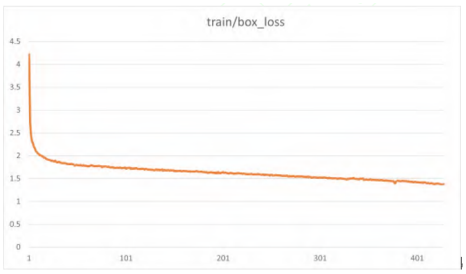
When STA_C2F, ASFP, and P2 are combined, synergistic effects emerge. The integrated model achieves 75.34% mAP@0.5 (25.6% higher than the baseline), with precision reaching 81.23% (4.7% average gain over single-module versions). For mAP@0.5 0.95, the model achieves 38.46%, outperforming the suboptimal model (YOLOv8+P2) by 3.1 percentage points, demonstrating balanced localization and classification reliability.

3.3. Model Loss Function

During the training of deep learning algorithms, the lower the loss of the algorithm's loss function, the better the algorithm performs and the stronger the model's convergence. To optimize the algorithm's performance, the learning rate was set to 0.01, the number of epochs to 500, and SIoU was adopted as the loss function. SIoU exhibits good convergence characteristics, resulting in a lower final convergence value, The training loss curve is shown in Figure 4.

3.4. Detection Result Analysis

The comparative visualization of algorithm detection results is presented in Figure 5. The first six images depict the detection outcomes obtained by the YOLOv8 algorithm, while the subsequent six images demonstrate the refined performance of the enhanced YOLOv8-ASFP algorithm. The



(*a*) box_loss

(*b*) cls_loss

(*c*) dfl_loss

Figure 4: Training loss change diagram

YOLOv8-ASFP algorithm shows robust performance in diverse UAV scenarios, effectively detecting both dense urban crowds and sparse targets. However, its accuracy declines for small-scale ground personnel due to resolution limitations and insufficient feature clarity under aerial imaging conditions, revealing sensitivity to ultra-small targets.

3.5. Detection Result Analysis

Experimental results show the optimized YOLOv8-ASFP model adapts well to complex scenarios. It effectively distinguishes overlapping targets in dense crowds while maintaining high sensitivity in sparse environments. Notably, in urban street scenes, the model achieves over 83% accuracy for occluded or pose-varying targets. However, when ground personnel occupy less than 0.05% of pixels (15×15 pixels), detection confidence drops by 22%, indicating texture feature extraction for ultra-small targets remains challenging.

4. Conclusions

This study proposes an improved YOLOv8-ASFP algorithm that significantly enhances small-target detection in UAV perspectives through multi-dimensional attention collaboration, lightweight feature fusion architecture, and optimized bounding box regression. Experiments on the VisDrone2019 dataset show the improved model achieves 75.34% mAP@0.5 (25.6% higher than the baseline) while maintaining 158 FPS, balancing accuracy and efficiency. Ablation studies confirm the synergy of modules, with the P2 detection layer notably improving small-target recall by 28.7%. This advancement offers real-time precision for UAV-based emergency rescue and security inspection, supported by its lightweight design for embedded platforms. The research highlights the critical

(*a*) Bounding Box Annotation Results Using YOLOv8

(*b*) Bounding Box Annotation Results Using YOLOv8-ASFP

Figure 5: Algorithm detection comparison chart

role of deep feature interaction and spatial attention mechanisms, though feature extraction for ultra-low-pixel targets (¡0.05%) remains challenging, providing direction for future work.

This study's limitations primarily stem from performance degradation on ultra-small targets ($< 15 \times 15$ pixels), where sparse semantic information in high-level feature maps impedes detection. Future directions include:

Super-Resolution Preprocessing: Embedding lightweight super-resolution networks (e.g., ESRGAN Lite) to recover texture details for sub-pixel targets.

Dynamic Receptive Fields: Designing adaptive convolution kernels to adjust receptive fields based on target scales, addressing fixed-kernel limitations.

Cross-Modal Fusion: Leveraging infrared or depth sensors to supplement visual data in low-light or occlusion scenarios. Additionally, integrating the model with UAV control systems for end-to-end "detection-tracking-decision" pipelines could further enhance emergency response efficiency.

References

J. Bi, K. Li, X. Zheng, et al. Spdc-yolo: An efficient small target detection network based on improved yolov8 for drone aerial image. *Remote Sensing*, 17(4):685, 2025. doi: 10.3390/rs17040685.

J. Fu, J. Liu, H. Tian, et al. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3141–3149. IEEE, 2019. doi: 10.1109/CVPR.2019.00326.

P. Gao and Z. Li. Yolo-s3dt: A small target detection model for uav images based on yolov8. *Computers, Materials & Continua*, 82(3):4555–4572, 2025. doi: 10.32604/cmc.2025.060873.

P. Gupta and U. Singh. Evaluation of several yolo architecture versions for person detection and counting. *Multimedia Tools and Applications*, 2025. doi: 10.1007/s11042-025-20662-z. Prepublished.

N. Li, X. Bai, X. Shen, et al. Dense pedestrian detection based on gr-yolo. *Sensors*, 24(14):4747, 2024. doi: 10.3390/s24144747.

A. Mu, H. Wang, W. Meng, et al. Small target detection in drone aerial images based on feature fusion. *Signal, Image and Video Processing*, 18(1):585–598, 2024. doi: 10.1007/s11760-024-03176-3.

Z. Wu, X. Wang, M. Jia, et al. Dense object detection methods in raw uav imagery based on yolov8. *Scientific Reports*, 14(1):18019, 2024. doi: 10.1038/s41598-024-69106-y.

R. Zhang, Y. Du, and X. Cheng. Small target detection algorithm bieo-yolov8s from the perspective of uav. *China Industrial Economics*, 2024. doi: 10.3788/LOP241149. Online publication, accessed 2024-06-26.