

TDF-YOLOv8 : An Enhanced UAV Object Detection Method Based on Triplet Attention and Dynamic Scaling

Bo Tan^{1, a}, Yuyong Cui^{1, b *}

¹Southwest Institute of Technical Physics, Chengdu 610041, China

^a 1281562824@qq.com, ^b 44989183@qq.com

Abstract: In this paper, challenges in object detection based on UAV imagery are addressed. The focus is on high-resolution imagery, varied perspectives, and dense arrangements of small objects. The YOLOv8s model is enhanced by integrating Triplet Attention into its C2f module to improve recognition precision of densely distributed targets. A novel Detect_Dyhead structure is introduced to dynamically adjust detection strategies across different scales and shapes. Additionally, the FocalerShapeloU loss function is employed to refine bounding box accuracy. Experimental results on the VisDrone2019 dataset show that Precision, Recall, mAP@0.5, and mAP@0.5:0.95 metrics are enhanced by 2.1%, 1.4%, 1.4%, and 1.2%, respectively, while the model size is reduced by 0.5MB. Strong performance is also demonstrated on the WiderPerson and custom Tanks datasets, underscoring the generalizability of the proposed TDF-YOLOv8 model.

Keywords: YOLOv8s, UAV aerial images, target detection, feature fusion.

1. Introduction

With the rapid development of UAV technology, the application of UAVs in fields such as environmental monitoring, disaster assessment, urban management, and security surveillance has become increasingly widespread. Target detection technology based on UAV aerial images has attracted much attention due to its great potential in these areas. However, traditional target detection algorithms often face significant challenges when dealing with UAV aerial images characterized by high resolution, variable view angles, and small, densely packed targets.

The YOLO (You Only Look Once) series of algorithms, as an end-to-end target detection method, has been widely used in target detection due to its advantages of fast detection speed and high accuracy. In January 2023, the Ultralytics team released YOLOv8, which improves detection accuracy and processing speed compared to its predecessor. Innovations such as Mosaic data enhancement, the C2f module, and anchor-free detection strategies make YOLOv8 particularly effective in complex scenes.

In the field of target detection for UAV aerial images, various improvement schemes have been proposed. For example, Liu et al. introduced an improved variant of YOLOv8s using Special Feature Pyramid Networks (SFPNs) and replaced some original components with more efficient ones while pruning the model to reduce redundancy[1]. Saydirasulovich's team utilized the BiFormer attentional mechanism and a novel IoU loss function to develop a better smoke detection method for UAVs[2]. Gang Wang and colleagues designed the Focal FasterNet block (FFNB) to achieve deeper feature integration in their UAV-YOLOv8 model and optimized it for aerial scene object detection[3].

Despite these improvements, challenges such as pixel sparsity, dense target distribution, and category imbalance remain in UAV aerial image target detection. Therefore, this paper proposes an improved YOLOv8s network structure aimed at enhancing the accuracy and robustness of target detection in UAV aerial images. The main contributions of this paper include: introducing the Triplet Attention mechanism in the C2f module to improve accuracy in handling densely distributed targets; improving the detection head to the Detect_Dyhead structure to enhance recognition of targets with different scales and shapes; and using the FocalerShapeIoU loss function to further optimize bounding box regression accuracy.

Comprehensive experimental results show that the improved model significantly enhances the accuracy and robustness of target detection, especially for densely distributed small targets in UAV aerial images. Specifically, the improved model demonstrates higher precision, recall, and mean average precision (mAP) under different IoU thresholds.

The rest of the paper is organized as follows: Section II details the architecture and key components of the YOLOv8 model; Section III describes the proposed improved YOLOv8 network architecture for UAV aerial image target detection; Section IV presents the experimental design and result analysis; and Section V concludes the paper and discusses future research directions.

2. YOLOv8 Algorithm

The YOLO (You Only Look Once) series of algorithms is renowned for balancing speed and accuracy, making it ideal for fast and accurate target identification in various scenarios. YOLOv8, introduced in January 2023 by Ultralytics, continues this tradition with five versions (n, s, m, l, and x) to meet diverse needs. YOLOv8s, in particular, is chosen in this paper for its balance between performance and hardware resource consumption, demonstrating exceptional effectiveness for target detection in UAV aerial images[4].

Built on YOLOv5, YOLOv8 has been comprehensively upgraded. Mosaic data enhancement, MixUp, and CopyPaste technologies have been introduced to increase training set diversity and model adaptability to complex scenes. Deep features are effectively extracted, and the model is made lightweight by combining the C2f module and SPPF structure in the Backbone. The multi-scale feature fusion is optimized by merging FPN and PAN concepts in the Neck design, enhancing precision and recall in multi-level target detection. Classification and localization tasks are independently handled by the decoupled output head, reducing interference. Bounding box prediction is simplified by the anchorless detection strategy, improving speed and reducing overfitting. YOLOv8's performance is optimized by employing dynamic sample allocation and a combined strategy of VFLLoss, DFLLoss, and CIOULoss in the loss function[5].

YOLOv8's processing speed, resource efficiency, and detection accuracy have been enhanced by these innovative designs, providing top-notch target detection solutions for mobile devices, drones, and other platforms.

3. Improved YOLOv8 Network Architecture

Aiming at the unique challenges of target detection in UAV aerial images, such as pixel sparsity, dense distribution, and category imbalance, an improved target detection algorithm, TDF-YOLOv8, has been designed using YOLOv8s as the baseline model. The specific improvements are as follows: the Triplet Attention mechanism has been introduced in the C2f module, which applies attention in the time, frequency, and channel dimensions synchronously, allowing key features to be efficiently screened and enhanced, thereby improving the model's accuracy in handling densely distributed targets; the detection head has been improved to the Detect_Dyhead structure with dynamic convolution technology, allowing the model to adaptively adjust detection strategies based on the actual size and shape of the target, which enhances the recognition ability for targets of different scales and shapes, making the model better suited to the complexities of UAV aerial image target detection; and the original CIOU loss function has been replaced with the FocalerShapeIoU loss function to further improve detection accuracy. These improvements enable TDF-YOLOv8 to better address the challenges of target detection in UAV aerial images.

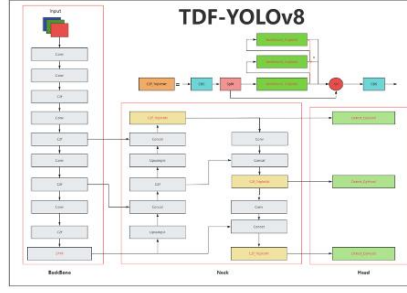


Fig. 1 TDF-YOLOv8 network structure

3.1 Incorporating the Triplet Attention mechanism into the C2f module

Triplet Attention proposes a novel self-attention layer design that integrates three different dimensions of attention mechanisms—Temporal, Spatial, and Channel—to comprehensively capture complex patterns in sequence data. A three-branch structure is utilized to facilitate cross-dimensional interactions within input data for computing attention weights, thereby establishing interdependencies between input channels or spatial locations at low computational cost. This unique approach analyzes input data from three distinct perspectives, akin to three observers viewing the same painting from different angles and collectively deciding on how to interpret the information. Among the three branches, attention weights for channel dimension C and spatial dimension W are computed by the upper branch through Z-pooling on the input tensor, followed by convolutional and Sigmoid operations to generate attention weights. The middle branch captures dependencies between channel dimension C and spatial dimensions H and W , using similar Z-pooling and convolution operations followed by Sigmoid activation. Dependencies between spatial dimensions while preserving input identity are captured by the lower branch, which conducts Z-pooling and convolution operations and generates attention weights via Sigmoid. After generating attention weights (Permutation), inputs are aligned by each branch, and their outputs are averaged (Avg) to yield the final Triplet Attention output. Information across different dimensions is synthesized by this structure, enhancing the model's ability to capture intrinsic data characteristics through rotation and alignment operations, and computational efficiency is offered, making it suitable for integration as a module in existing network architectures to handle complex data structures effectively[6].

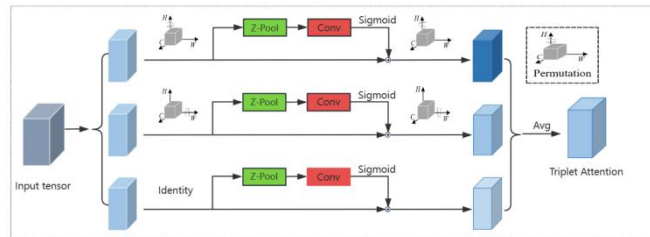


Fig. 2 Triplet Attention Schematic

Triplet Attention has been incorporated into the Bottleneck of the C2f module to form Bottleneck_TripletAt, and all the Bottleneck instances have been replaced with Bottleneck_TripletAt to form the improved C2f_TripletAt. By fusing Triplet Attention, lightweighting has been achieved, and additional performance improvements have been gained.

3.2 Improve detection head Detect to Detect_Dyhead

The core idea of the Dynamic Head structure is to have the network's structure or weight assignment adaptively adjusted based on input data features, enhancing its ability to handle diverse classes or scenarios effectively. In target detection, the detection head is dynamically modified based on the scale, shape, or position of target objects, improving detection accuracy across various scenarios[7].

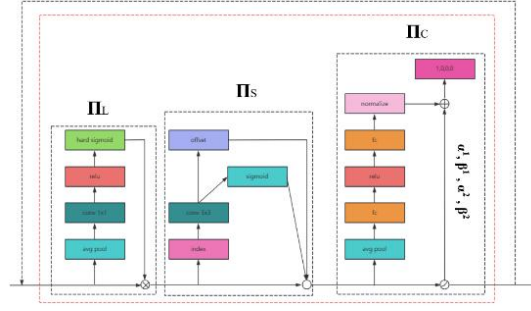


Fig. 3 DyHead Block Schematic

As shown in Fig. 3, the DyHead module schematic illustrates three key attention mechanisms: scale-aware attention (Π_L), spatial-aware attention (Π_S), and task-aware attention (Π_C). Each mechanism is focused on different aspects to refine feature maps, thereby improving clarity and focus. Features are organized into a 3D tensor $F \in \mathbb{R}^{L \times S \times C}$ at the same scale by scale-aware attention, features are aggregated based on spatial locations by spatial-aware attention, and feature channels are dynamically adjusted to adapt to various detection tasks, such as classification and regression, by task-aware attention.

Initially, feature maps often exhibit noise due to domain variations. However, the application of these attention mechanisms refines the feature representation step-by-step, optimizing it for different target detection tasks.

3.3 Improve the loss function to FocalerShapeIoU

Focaler-IoU and Shape-IoU are advanced bounding box regression methods. Focaler-IoU enhances detection performance by adjusting IoU mapping linearly, focusing on regression samples of varying difficulty[8]. Shape-IoU improves accuracy by integrating shape and scale factors, considering box geometry in loss calculation[9]. FocalerShapeIoU combines these approaches, optimizing regression by adjusting IoU and adding shape-related terms, effectively handling sample imbalance and box geometry. Formula: IoU is Intersection over Union, and d and u are adjustable. Terms $distance_{shape}$ and Ω_{shape} relate to box shape and scale.

$$IoU_{focaler} = \begin{cases} 0 & \text{if } IoU < d \\ \frac{IoU - d}{u - d} & \text{if } d \leq IoU \leq u \\ 1 & \text{if } IoU > u \end{cases} \quad (1)$$

$$L_{FocalerShapeIoU} = 1 - IoU_{focaler} + distance_{shape} + 0.5 \times \Omega_{shape} \quad (2)$$

4. Experiment Design and Result Analysis

4.1 Dataset and Experimental Environment

The VisDrone2019 dataset was used in the experiments, which were collected and released by the AISKEYEYE team from Tianjin University's Machine Learning and Data Mining Laboratory. This dataset includes 288 video clips totaling 261,908 frames and 10,209 static images, captured by drone cameras across 14 cities in China, covering thousands of kilometers. It comprises 10 object categories: Pedestrian, People, Bicycle, Car, Van, Truck, Tricycle, Awning-Tricycle, Bus, and Motor. The dataset is divided into training (6,471 images), validation (548 images), and test sets (1,610 images). Most objects in the dataset are small and numerous.

Experiments were conducted using an NVIDIA GeForce RTX 3090 GPU (24GB), a 15-core CPU, PyTorch framework version 2.0, Python version 3.9.0, and CUDA version 11.8. Specific parameters of the experimental environment are detailed in Table 1.

Table 1. Traning parameter setting

Parameter	Setting	Parameter	Setting
epochs	300	close_mosaic	10

patience	50	Warmup_epochs	3.0
batch	16	lrf	0.01
imgsz	640	lr0	0.01
workers	8	momentum	0.937
optimizer	auto	Weight_decay	0.0005

4.2 Ablation experiments

The YOLOv8s model was enhanced, and the outcomes of each improvement were statistically analyzed, as depicted in Table 2. T, D, and F in the table denote the integration of Triplet Attention, Dynamic Head, and FocalerShapeIoU into the YOLOv8s model, respectively. The improved model proposed in this paper is referred to as TDF-YOLOv8s.

Table 2. The VisDrone2019 dataset ablation experiments

Algorithm	T	D	F	Size /MB	GFLOPs	P	R	mAP@0.5	mAP@0.5:0.95
YOLOv8s				22.5	28.5	0.505	0.381	0.392	0.232
T-YOLOv8	√			21.5	22.6	0.520	0.381	0.394	0.235
D-YOLOv8		√		20.9	28.1	0.521	0.391	0.399	0.24
TD-YOLOv8	√	√		22.0	28.3	0.51	0.391	0.4	0.24
TDF-YOLOv8	√	√	√	22.0	28.3	0.526	0.395	0.406	0.244

As shown in Table 2, varying degrees of enhancement in detection performance were achieved by the proposed improvements to YOLOv8s.

The integration of the Triplet Attention mechanism into the C2f module in T-YOLOv8 significantly reduced the number of floating-point operations and model size, while improving detection accuracy compared to YOLOv8s. Improvement 1 (T) was validated by TD-YOLOv8, which improved mAP@0.5 by 0.1 percentage points over D-YOLOv8.

D-YOLOv8, which replaced the Detect head with Detect_Dyhead similar to T-YOLOv8, reduced floating-point operations and model size while enhancing detection accuracy. Improvement 2 (D) was validated by TD-YOLOv8, which increased R by 1 percentage point, mAP@0.5 by 0.6 percentage points, and mAP@0.5:0.95 by 0.5 percentage points compared to T-YOLOv8.

Improvement 3 (F) was validated by TDF-YOLOv8, which modified the loss function to FocalerShapeIoU, improving P by 1.6 percentage points, R by 0.4 percentage points, mAP@0.5 by 0.6 percentage points, and mAP@0.5:0.95 by 0.4 percentage points compared to TD-YOLOv8.

TDF-YOLOv8 integrated all improvements, enhancing P, R, mAP@0.5, and mAP@0.5:0.95 by 2.1, 1.4, 1.4, and 1.2 percentage points respectively, while reducing the model size by 0.5MB. The significant enhancement of the improved model for drone-based aerial target detection is thus validated.

4.3 Comparative Experiments

To validate the superiority of the proposed improved model in drone-based aerial image target detection, comparative experiments were conducted between TDF-YOLOv8 and mainstream algorithms YOLOv5s, YOLOv5m, and YOLOv7-tiny on the VisDrone2019 dataset. The experimental results are shown in Table 3.

Table 3. The VisDrone2019 dataset comparative experiments

Algorithm	Size /MB	GFLOPs	P	R	mAP@0.5	mAP@0.5:0.95
YOLOv5s	14.4	15.8	0.474	0.344	0.343	0.189
YOLOv5m	42.2	48.0	0.514	0.37	0.378	0.219
YOLOv7-tiny	12.3	13.1	0.468	0.38	0.357	0.186
YOLOv8s	22.5	28.5	0.505	0.381	0.392	0.232
TDF-YOLOv8	22.0	28.3	0.526	0.395	0.406	0.244

As shown in Table 3, despite the lower parameter counts and computational loads of YOLOv5s and YOLOv7-tiny compared to the improved model proposed in this paper, their detection accuracy

is significantly inferior. The superiority of the proposed model in drone-based aerial image target detection is demonstrated, considering that the size of the improved model in this paper already meets the hardware constraints of most drone devices. Moreover, YOLOv5m and YOLOv8s exhibit lower P, R, mAP@0.5, and mAP@0.5:0.95 compared to the proposed algorithm, while their parameter counts and computational loads are considerably higher, further affirming the superior performance of the improved model.

4.4 Generalization Experiment Comparison

To verify that the proposed algorithm performs well on other drone-based aerial image datasets and exhibits sufficient generalization capability, experiments were conducted on the WiderPerson dataset and a custom Tanks dataset. The detection performance of the improved model TDF-YOLOv8 was compared against YOLOv8s. The results are presented in Table 4.

Table 4. Generalization experiment verification on WiderPerson and Tanks datasets

Dataset	Algorithm	Size /MB	GFLOPs	P	R	mAP@0.5	mAP@0.5:0.95
WiderPerson	YOLOv8s	22.5	28.4	0.792	0.655	0.758	0.475
	TDF-YOLOv8	22.0	28.3	0.791	0.659	0.761	0.478
Tanks	YOLOv8s	22.5	28.5	0.824	0.872	0.914	0.596
	TDF-YOLOv8	22.0	28.3	0.898	0.818	0.917	0.618

From Table 4, it can be observed that the improved model TDF-YOLOv8 in this paper has a slightly smaller size and computational load compared to YOLOv8s. However, its detection accuracy on the WiderPerson and Tanks datasets is significantly better than that of YOLOv8s, confirming the generalization capability of the proposed improved model.

5. Summary

In this paper, the performance of the YOLOv8s model in drone-based aerial image target detection is enhanced through targeted improvements, especially in complex scenarios like high resolution, diverse viewpoints, small targets, and dense distributions. The TDF-YOLOv8 model has Triplet Attention integrated into the C2f module, improving feature selection and reinforcing key features for more accurate dense object detection. The adoption of Detect_Dyhead dynamically adjusts detection strategies, significantly enhancing recognition of diverse target scales and shapes inherent to drone-captured images. The innovative FocalerShapeIoU loss function is employed to optimize bounding box regression, further improving detection precision.

In generalization experiments, it is shown that TDF-YOLOv8 outperforms the baseline YOLOv8s on WiderPerson and custom Tanks datasets, affirming robust generalization. Besides reducing size and computational load, notable enhancements in detection accuracy are achieved, advancing practical applications in environmental monitoring, disaster assessment, urban management, and security surveillance with drone-based target detection.

Future research should prioritize efficiency in extreme viewpoints, complex backgrounds, and dynamic scenes, integrating deep learning with traditional methods and advanced attention mechanisms to address class imbalance in small target detection. As drone technology and applications expand, continuous optimization of detection algorithms remains critical for enhanced intelligence and adaptability.

References

- [1] Z. Liu, "Unmanned Aerial Vehicles General Aerial Person-Vehicle Recognition Based on Improved YOLOv8s Algorithm," *Computers, Materials & Continua*, vol. 78, no. 3, Mar. 2024.
- [2] S. N. Saydurasulovich et al., "An improved wildfire smoke detection based on YOLOv8 and UAV images," *Sensors*, vol. 23, no. 20, pp. 8374, Oct. 10, 2023.
- [3] G. Wang et al., "UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios," *Sensors*, vol. 23, no. 16, pp. 7190, Aug. 15, 2023.

- [4] M. Hussain, "YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection," *Machines*, vol. 11, no. 7, pp. 677, Jun. 23, 2023.
- [5] C. H. Kang and S. Y. Kim, "Real-time object detection and segmentation technology: an analysis of the YOLO algorithm," *JMST Advances*, vol. 5, no. 2, pp. 69-76, Sep. 2023.
- [6] D. Misra et al., "Rotate to attend: Convolutional triplet attention module," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3139-3148.
- [7] X. Dai et al., "Dynamic head: Unifying object detection heads with attentions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7373-7382.
- [8] H. Zhang and S. Zhang, "Focaler-IoU: More Focused Intersection over Union Loss," *arXiv preprint arXiv:2401.10525*, Jan. 19, 2024.
- [9] H. Zhang and S. Zhang, "Shape-iou: More accurate metric considering bounding box shape and scale," *arXiv preprint arXiv:2312.17663*, Dec. 29, 2023.