BDS-YOLOv8: An Enhanced UAV-Based Algorithm for Small Object Detection

Kai Peng, Yang Xu

Abstract—**Many small targets in UAV aerial photography are affected by occlusion and lighting conditions. At present, the accuracy of some popular detection models is low, and there are missing and false detection phenomena. To solve the above problems, we optimized YOLOv8s and presented an advanced small target detection algorithm for UAV aerial photography (BDS-YOLOv8). First, add a tiny detection head and delete the original large detection head. To save the parameter cost of the model and realize a complete feature fusion process, the idea of a bidirectional feature pyramid network (BiFPN) is introduced into the neck network and combined with the detection head structure proposed in this paper. Second, the dynamic snake convolutional layer (DSConv) replaces the convolutional layer in the bottleneck in the backbone network to enhance the ability of the model to extract the feature information of small targets. Finally, Wise-IoU (WIoU) v3 was employed as the bounding box regression loss, and the influence of deviation on the loss was dynamically adjusted by introducing weight factors. The model was experimented on using the VisDrone2019 dataset. The experimental results show that compared with the baseline model YOLOv8s, the number of model parameters is reduced by 17.5%, and the average detection accuracy of mAP0.5 and mAP0.5:0.95 is increased by 5.6% and 3.7%, respectively.**

Index Terms—**UAV, Small Target Detection, DSConv, BiFPN**

I. Introduction

UAVs have been extensively applied in various fields, including urban security [1], traffic monitoring [2], crop analysis [3], and many other fields. With the popularization of UAV image acquisition, due to factors such as complex backgrounds, lighting conditions, high proportion of small targets, and occlusion, the recognition of small targets in UAV aerial photography has also become a corresponding problem. Traditional object detection algorithms have difficulty ensuring high-precision detection results when processing these images, as exemplified by the Deformable Part Model (DPM) [4], which utilizes a trained classifier to classify the local image in each window and produces a binary classification result. This detection method uses a pixel-by-pixel and window-by-window calculation method, which is very time-consuming and usually has low detection accuracy. Therefore, it is gradually being replaced by some existing mainstream algorithms.

Currently, mainstream algorithms for object detection can be classified into two-stage and single-stage categories. For the two-stage algorithm, such as Faster R-CNN [5], a Region Proposal Network (RPN) is introduced based on generating candidate regions, and the candidate region generation process is integrated with the object detection network. It dramatically improves the efficiency of detection. Although the two-stage algorithm offers certain benefits in detection accuracy, it also has shortcomings, such as difficult training and optimization. Different from the two-stage algorithm, the single-stage algorithm (such as YOLO [6] and SSD [7]) eliminates the traditional candidate region generation step and simultaneously performs object classification and bounding box regression in one forward propagation, which has higher computational efficiency and stronger real-time performance, but has limited accuracy for small targets.

Based on the above problems, this paper proposes the BDS-YOLOv8 object detection model, aimed at enhancing the detection performance for small targets. The key contributions are summarized as follows:

1. We introduced an innovative detection head structure that includes a tiny detection head while omitting the large detection head. To reduce the model’s parameter cost and achieve a complete feature fusion process, the idea of a bidirectional feature pyramid network (BiFPN) is used in the neck network to combine the detection head structure proposed in this paper.
2. We use Dynamic Snake Convolution (DSConv) to replace conventional convolution layers in Bottleneck, enhancing the model’s ability to extract feature information of small objects under challenging conditions.
3. We use Wise-IoU (WIoU) v3 as the model’s bounding box loss function. It incorporates a dynamic non-monotony mechanism and an improved method for gradient gain allocation. Introducing a weighting mechanism solves traditional IoU’s limitations in dealing with small targets, multiple targets, and complex scenarios.
4. We performed relevant experiments based on the above points and visually analyzed the results.

II. Related Work

Target detection from the UAV perspective is widely used but also faces many challenges. Because of its small targets, target occlusion, and target clustering, it has important practical value and research significance [8]. With the continuous development of object detection technology, many excellent methods have emerged to solve the above problems. Yang et al. [9] proposed a QueryDet detection algorithm, which uses a novel query mechanism cascaded

Manuscript received March 12, 2025; revised May 17, 2025.

This research was funded by the project of the Liaoning Provincial Department of Education (Project Number: LJKZ0310)

Kai Peng is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 2415784248@qq.com)

Yang Xu is a professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author to provide phone: +8613889785726; e-mail: xuyang_1981@aliyun.com)

sparse query (CSQ) to accelerate the inference of dense object detectors based on feature pyramids. The region containing small objects is initially screened by both high and low-resolution features. To some extent, it solves the problem of poor detection effects of small objects. Yang et al. [10] introduced the ClusDet framework, which consists of three components: the Clustering Proposal Network (CPNet), the Scale Estimation Network (ScaleNet), and the Detection Network (DetecNet). CPNet identifies clustered object regions from the input image, while ScaleNet predicts their corresponding object scales. These normalized regions are then passed to DetecNet to perform the final detection, which solves the problem of small object clustering to a certain extent. Xu et al. [11] proposed an occlusion object detection algorithm, which improved the detection performance of occlusion scenes through the combination of the occlusion sample generation module (OSGM) and the occlusion sample inpainting module (OSIM).

Although the current mainstream object detection methods have made remarkable progress in promoting the small target detection task of UAVs [12], they are usually accompanied by high computational requirements. They are difficult to apply to resource-constrained low-power image processors [13]. Therefore, it is a great challenge to achieve the balance between computing requirements and detection effects, and the YOLO series of detection algorithms provides an effective solution. In 2016, Joseph Redmon et al. presented YOLOv1, the first version of YOLO series. As a single-stage object detection algorithm, YOLOv1 provides a new idea for object detection. Unlike conventional detectors, it transforms object detection into a regression problem, which adopts multi-stage processing. The model extracts bounding boxes and estimates class probabilities based on the image content. Subsequently, Joseph Redmon et al. released YOLOv3 in 2018, which uses three-level feature maps (small, medium, and large) for multi-scale detection and introduces a K-means clustering algorithm to optimize the size of anchor frame and optimizes the detection performance of small objects. In 2020, Ultralytics released YOLOv5, it widely improved in the YOLO family and offered models in different sizes (e.g., YOLOv5s, v5m, v5l, v5x) to adapt to different hardware platforms, providing researchers with more flexible options for their mission needs. In 2022, Meituan Vision AI released YOLOv6, which uses a lightweight backbone incorporating CSPStackRep structures and a PAN topological neck to further enhance feature extraction capabilities. In 2023, Ultralytics released YOLOv8 based on YOLOv5; as a cutting-edge variant within the YOLO family, it abandoned the previous Anchor-Base and used the Anchor-Free idea. This change means the model no longer relies on predefined anchor boxes during object detection but directly predicts the bounding boxes and categories of objects, thus reducing the complexity of the model structure and enhancing the flexibility and efficiency of detection.

This paper chooses YOLOv8s as the baseline model, consisting of backbone, neck, and head networks. The basic network structure of YOLOv8s is presented in Fig. 1. It first extracts features from input images through the backbone network to generate feature maps of different scales. Then, these feature maps are fused through the neck network (FPN [14]-PAN [15]) to enhance the feature expression and multi-scale object detection. Finally, the head network makes the prediction, and the object's bounding box and class confidence are output.

III. Improved Model

Building on the YOLOv8s architecture, we designed the BDS-YOLOv8 model. The overall architecture of the model is presented in Fig. 2. In the design of the detection head of the model, a tiny detection head is added, which removes the original large detection head, decreases the number of model parameters to a certain degree, and increased the model's attention to small targets. In the backbone network structure of the model, the C2f structure is improved, and the conventional convolution layer in Bottleneck is replaced with a dynamic snake convolution layer, which boosts the model's capability for small target detection in some complicated backgrounds to a certain extent. In the neck network structure of the model, the module PAN-FPN of YOLOv8 was

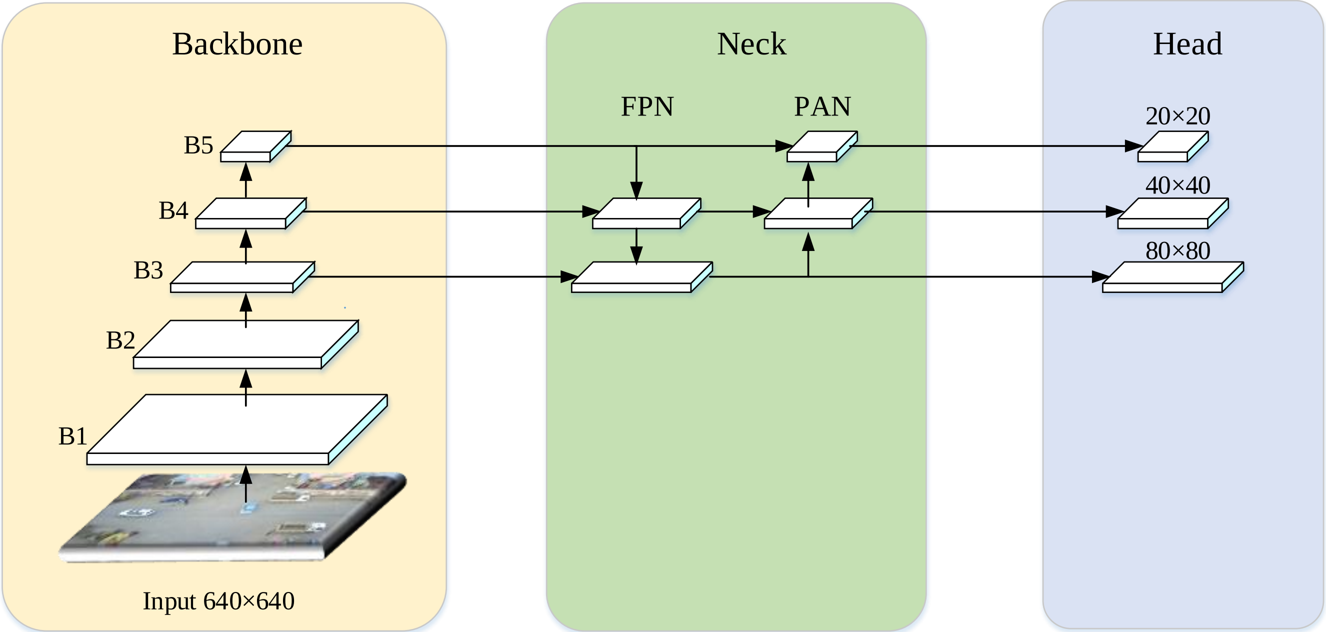


Fig. 1. YOLOv8 network structure

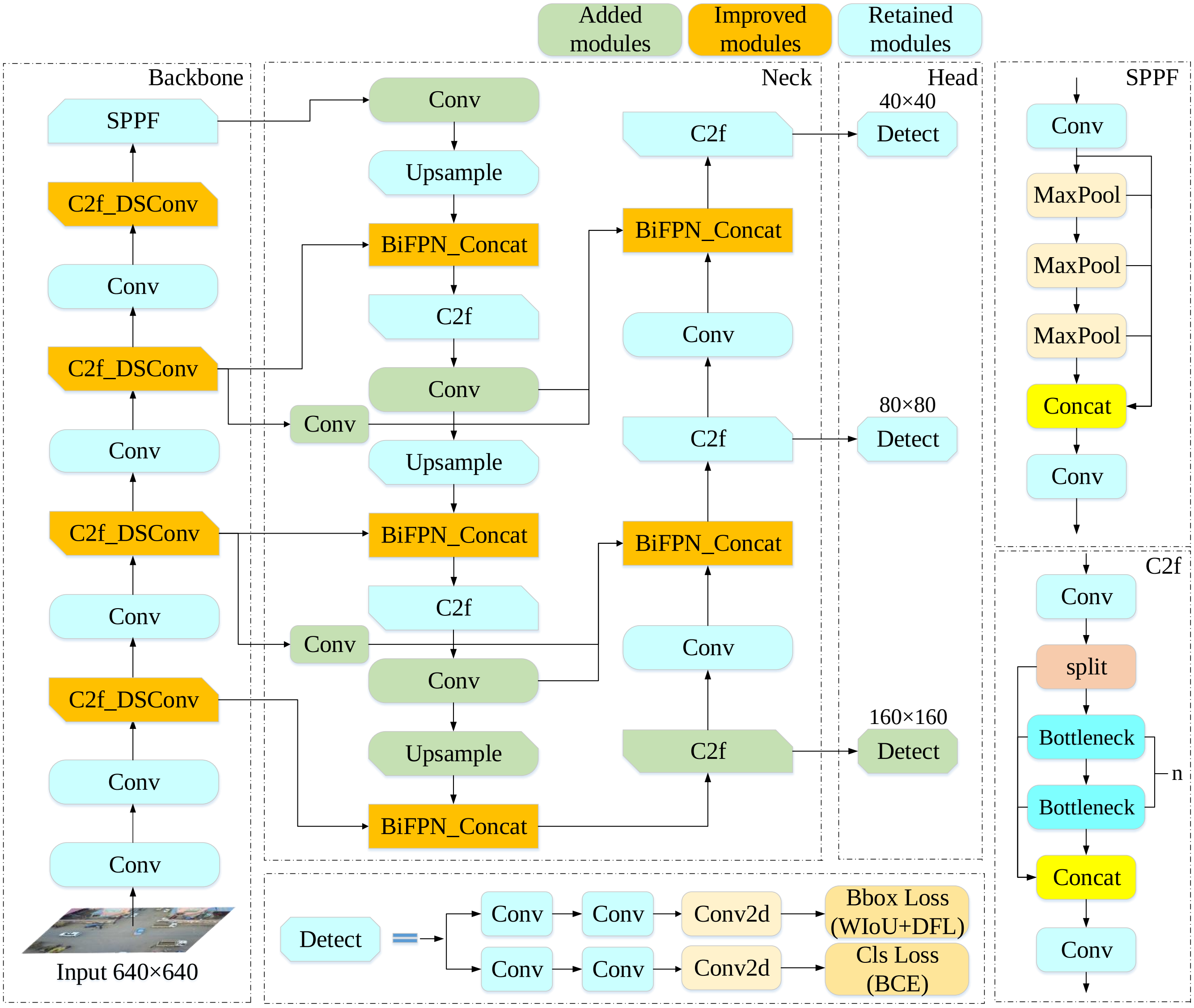


Fig. 2. BDS-YOLOv8 network model

improved and replaced with the module combining bidirectional feature pyramid network and path aggregation network (Bi-FPN-PAN), which effectively reduced the parameter cost of the model and realized a complete feature fusion process. In the model's loss function, WIoU v3 is used as the bounding box regression loss, and the influence of deviation on the loss is dynamically adjusted by introducing weight factors to optimize the model's location performance to the target bounding box.

A. Improved detection head

By default, YOLOv8 is equipped with detection heads of three different scales to detect targets. Specifically, the small detection head is associated with an 80×80 feature map and is responsible for identifying small objects larger than 8×8. The medium detection head corresponds to a 40×40 feature map, targeting objects exceeding 16×16 in size. Lastly, the large detection head operates on a 20×20 feature map and is used for detecting large-scale objects with dimensions greater than 32×32. Through experiments, we find that this head structure is ineffective in detecting small objects, so the original head structure is improved.

As illustrated in Fig. 3, the network architecture is modified by introducing a new tiny detection head and removing the original large detection head. The added tiny head is associated with a high-resolution feature map of size 160×160, which enables the detection of tiny objects larger than 4×4.

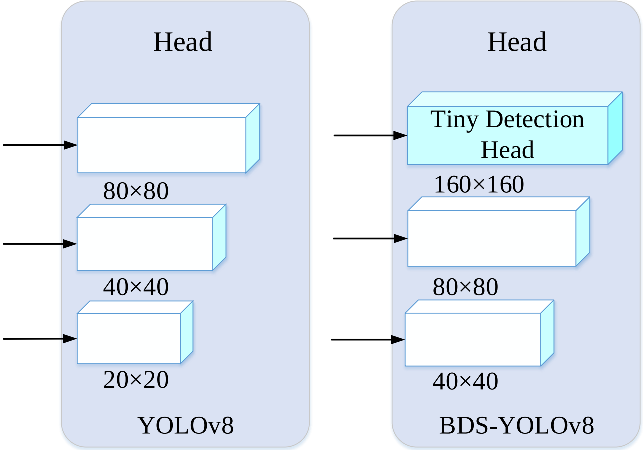


Fig. 3. Detection head comparison

B. Improved backbone

YOLOv8 combines the conventional convolution module and C2f module, wherein the conventional convolution module is employed to capture essential features, while the C2f module enhances multi-scale feature fusion capability through cross-layer connection and feature aggregation, thus strengthening the detection effectiveness of the model. A Bottleneck module in the traditional C2f structure uses a fixed convolution mode, but the traditional convolution operation may not capture the features of small targets adequately when dealing with complex backgrounds or textures, which tends to cause fuzzy features or information loss. Therefore, dynamic snake convolution [16] is used in this paper to replace conventional convolution layers in Bottleneck, improving the feature extraction effect and reducing redundant features, making the model more accurate while maintaining efficiency. The four different convolution kernels are shown in Fig. 4. (a) standard convolution is simple and efficient, (b) dilated convolution [17] is suitable for capturing context information, while (c) deformable convolution [18], and (d) dynamic snake convolution significantly enhance feature extraction capabilities for complex backgrounds, small targets and, irregular shapes by flexibly adjusting sampling locations. In particular, dynamic snake convolution performs best in boundary feature extraction and complex scenes.

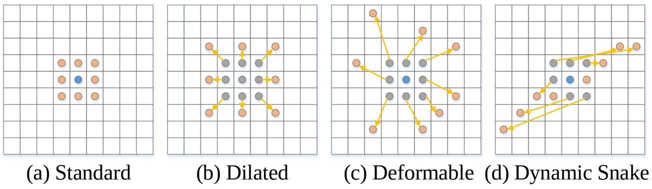

(a) Standard (b) Dilated (c) Deformable (d) Dynamic Snake

Fig. 4. Comparison of convolution kernel structure

The inspiration for dynamic snake convolution comes from observing and understanding the particularity of tubular structures. However, the UAV aerial photography data set used in this paper has a tubular structure, such as a road. Dynamic snake convolution has a good feature extraction ability for this tubular structure. Dynamic snake convolution accurately captures the features of tubular structures by adaptively focusing on elongated and circuitous local structures. The main concept behind this convolution method is to enhance the ability to perceive and optimize the feature extraction of tubular structures through the convolution kernel of dynamic shapes. We improve the C2f structure in the YOLOv8s architecture's backbone and replace the conventional convolution layer in Bottleneck with a dynamic snake convolution layer. The C2f_DSConv structure is shown in Fig. 5.

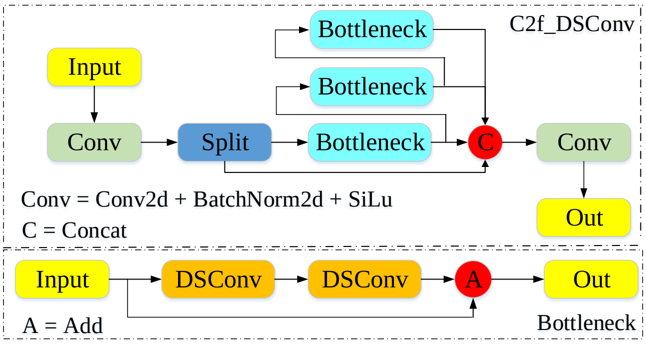


Fig. 5. C2f_DSConv structure

C. Improved neck

In YOLOv8, the feature graph is mainly divided into five scale features in order from largest to smallest, and they are processed through the FPN (F3–F5) and PAN (P4–P5) structures built upon the backbone (B1–B5) and neck networks. Initially, five multi-scale features (B1–B5) are extracted from the input image by the backbone. B1 is the large-scale feature map, which contains rich, detailed information and is suitable for small target detection. B5 is a small-scale feature map containing high-level semantic information suitable for large target detection. Secondly, these basic feature maps are further processed and optimized by FPN-PAN. FPN uses a top-down structure to improve semantic information in multi-scale feature blocks (B3-F3, B4-F4) and further improve the expression ability of low-level features. FPN-PAN adds a bottom-up structure based on FPN. Using bottom-up feature fusion strengthens the localization ability of mid-high-level feature maps (F4-P4, F5-P5). The position information lost in the semantic enhancement of FPN is made up, and the semantic and localization features are complemented.

However, the structure of FPN-PAN still has room for improvement when processing small target detection scenarios such as UAV aerial photography data. On the one hand, insufficient attention to large-scale features (such as B1

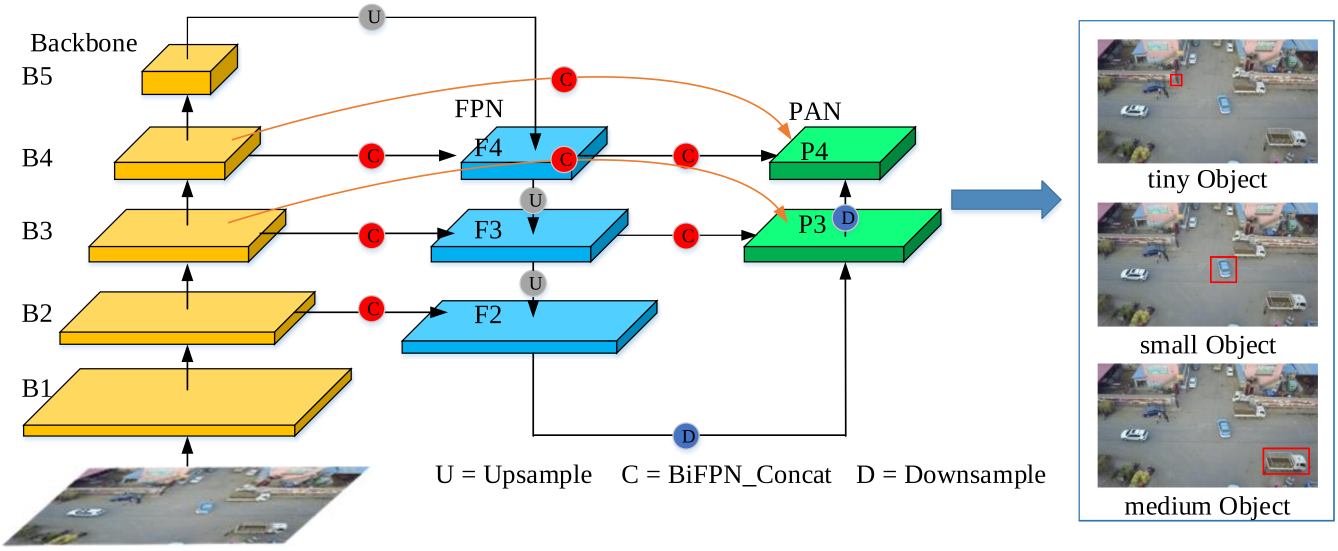


Fig. 6. Improved neck network

and B2) may lead to the degradation of small target detection performance. On the other hand, long-term up-sampling and down-sampling operations may lead to feature degradation or information loss. Accordingly, several modifications have been applied to the neck structure in this study.

First, to strengthen the extraction of large-scale semantic information (such as B1 and B2) in the traditional FPN-PAN structure, we added an upsampling process from F3 to F2 in FPN. We fused the features extracted from the B2 layer of the backbone network, enhancing the detection performance for small targets. Considering the model's efficiency, we deleted the downsampling process from P4 to P5 in PAN to reduce the model's parameters and improve the efficiency. To enrich the information content of low-resolution feature maps and improve final detection accuracy, an up-sampling layer and a 1×1 convolution layer were added at each stage of the neck. The up-sampling operation progressively enlarges the feature map dimensions, while the 1×1 convolution reduces the number of channels to retain computational efficiency. We call this structure the tiny target detection layer.

Secondly, the concept of BiFPN [19] is adopted to enhance the feature fusion strategy within the neck network. Its core objective is to increase the efficiency and frequency of multi-scale feature interaction, thereby optimizing the exchange of information across different resolutions. This facilitates better utilization of low-level detailed features and high-level semantic cues, ultimately leading to improved detection accuracy. The main realization principle is to maintain the feature fusion method of neck FPN and optimize the feature fusion method of neck PAN. In the PAN structure, when a feature map has two input paths and its resolution matches that of a corresponding feature map in the backbone network, an additional connection is introduced from the backbone to fuse this feature map.

Finally, each bidirectional (FPN and PAN) path is treated as a unit, and the unit is reused to enhance integration. To make the model lightweight, we added only additional paths of B3-P3 and B4-P4 and used only one cell. The enhanced neck structure is illustrated in Fig. 6.

D.Improved loss function

In target detection tasks under UAV aerial photography scenarios, the significant presence of small objects makes the design of a suitable loss function critical for enhancing the model's detection efficiency. The original YOLOv8 adopts Distribution Focal Loss (DFL) [20] and Complete Intersection over Union (CIoU) [21] to compute the bounding box regression loss. However, using CIoU has the following shortcomings: First, CIoU aims to optimize factors including the bounding box's aspect proportions and center point location, but for small targets with significant shape variations, CIoU may struggle to handle precise bounding box regression effectively. Second, CIoU does not consider the difficulty of regressing a target bounding box. It may excessively optimize easily regressed samples while failing to provide sufficient optimization for difficult-to-regress samples. Third, CIoU involves inverse trigonometric functions in its computation, which increases computational complexity and resource consumption and is not conducive to the lightweight of the model. The calculation formulas for CIoU are shown in Equations (1)–(3).

$$CIoU = IoU - \frac{\rho^2\left(b, b^{gt}\right)}{c^2} - av \tag{1}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{2}$$

$$\alpha = \frac{v}{\left(1 - IoU\right) + v} \tag{3}$$

In Equation (1), *IoU* denotes the ratio of the intersection area to the union area between the predicted and ground truth boxes; $\rho(b,b^{gt})$ denotes the Euclidean distance between the centroids of the predicted and ground truth boxes; *c* represents the diagonal length of the smallest enclosing box covering both. In Equation (2), *h* and *w* denote the predicted box's height and width, and h^{gt} and w^{gt} describe the height and width of the ground truth box. In Equation (3), *v* represents the variation in aspect ratios between the predicted and reference boxes, and *α* is a dynamic factor that modulates the relative importance of *IoU* and aspect ratio in the loss computation.

The previously described CIoU employs a static focusing mechanism. In contrast, WIoU comprehensively considers IoU, centroid offset, and aspect ratio error while providing an improved gradient gain distribution method, successfully boosting the model's effectiveness in complex scenes and small target detection. Tong et al. [22] presented WIoU, which includes various versions: WIoU v1 constructs an attention-based bounding box loss, and WIoU v2 and WIoU v3 further incorporate a focusing coefficient. This paper adopts the latest version, WIoU v3, which makes use of a dual-layer attention module and a dynamic non-monotony mechanism, further optimizing the computation of both the attention mechanism and the focusing coefficient to make them more dynamic and efficient. Its calculation formulas are shown in Equations (4)–(6).

$$L_{WIoUv3} = L_{IoU}\exp\left(\frac{\left(x_p - x_{gt}\right)^2 + \left(y_p - y_{gt}\right)^2}{\left(W_g{}^2 + H_g{}^2\right)^*}\right)\gamma \tag{4}$$

$$\gamma = \beta / \delta\alpha^{\beta-\alpha} \tag{5}$$

$$L_{IoU} = 1 - IoU \tag{6}$$

In Equation (4), L_{IoU} denotes the IoU-based loss, which evaluates the overlap between the predicted and the ground truth boxes. x_p and y_p represent the coordinates of the predicted box, while x_{gt} and y_{gt} represent the coordinates of the ground truth box; *H* and *W*, respectively, stand for the height and width of the two boxes. In Equation (5), *β* denotes the anomaly degree of the predicted box; the smaller the anomaly degree, the better the anchor box quality. Using *β* to construct a non-monotony focusing function assigns a smaller gradient gain to predicted boxes with larger anomaly values, reducing harmful gradients from low-quality training samples; α and δare hyperparameters. In Equation (6), *IoU* represents the intersection-over-union between the predicted box and the ground truth box, and the meanings of the other parameters are shown in Fig. 7.

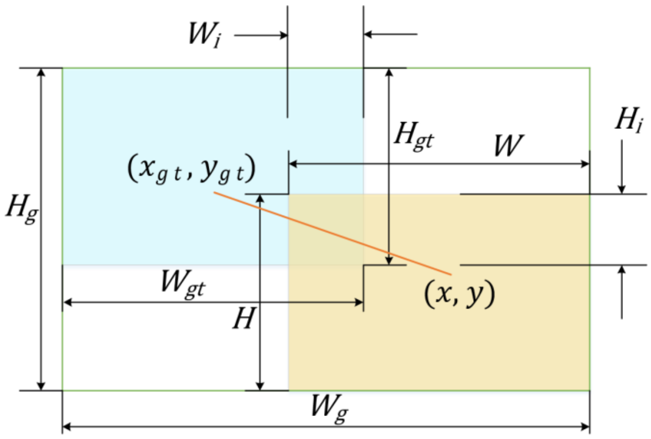


Fig. 7. WIoU v3

IV. Experiment And Analysis

A. Experimental Dataset

In this experiment, we use the VisDrone2019 dataset [23], a popular UAV aerial photography small target dataset gathered and organized by the AISKYEYE team at Tianjin University. The images in this dataset were captured from 14 cities in China, with a high resolution of up to 2000×1500 pixels. The training set consists of 6,471 images in the datasets, with 343,205 annotated targets, averaging 53 instances per image, indicating a high target density. Some targets are even smaller than 8×8 pixels. A total of 548 images are used for validation and 1,610 for testing in the datasets. The dataset covers 10 target types: pedestrian, person, bicycle, car, van, truck, tricycle, awning tricycle, bus, and motor vehicle. This dataset contains many small targets that are densely and unevenly distributed. VisDrone2019 is a comprehensive UAV dataset characterized by multi-scale, multi-scene, and multi-angle images compared to traditional computer vision datasets, making it more challenging than general computer vision tasks [24].

B. Experimental Equipment and Training Strategy

The model used in this experiment is YOLOv8. First, it is necessary to set up a compatible version of PyTorch. Additionally, considering the model's computational cost and the dataset's size, renting a suitable server is necessary. Table I shows the server's hardware specifications and the required software environment configurations, including the GPU model, memory, operating system, and deep learning framework.

TABLE I
Training Environment

Parameters	Configuration
GPU	NVIDIA GeForce RTX3090
GPU memory size	24 G
Operating systems	ubuntu22.04
Deep learning architecture	PyTorch2.1.2+Cuda11.8

The YOLOv8 framework enables the creation of five model variants based on scale: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. Users can balance model accuracy and inference speed based on task requirements. This paper focuses on small target detection in UAV aerial photography and considers speed and accuracy. Ultimately, we select the YOLOv8s as the baseline models of this study. YOLOv8s has low parameters and computational costs while maintaining good performance in small target detection.

During the training phase, essential hyperparameters, including batch size, number of epochs, and optimizer were meticulously adjusted. The detailed configurations are provided in Table II. These adjustments improve the model's detection performance while ensuring efficient inference speed.

TABLE II
Hyperparameters for Model Training

Parameters	Setup
Epochs	200
Input image size	640 × 640
Batch size	6
Momentum	0.932
Initial learning rate	0.01
Final learning rate	0.0001
Optimizer	SGD

C. Evaluation Metrics

This paper adopts precision, recall, mAP0.5, and mAP0.5:0.95 as evaluation metrics to evaluate the improved model's detection performance. These metrics are defined in detail below.

Precision quantifies the ratio of true positive predictions to the total number of instances classified as positive. This process is mathematically defined as:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall measures the model's ability to detect all relevant positive instances within the dataset. This process is mathematically defined as:

$$Recll = \frac{TP}{TP + FN} \tag{8}$$

Average Precision (AP) is obtained by computing the precision for each class based on ranked outputs and then taking the average across all classes. This process is mathematically defined as:

$$AP = \int_0^1 Precision(Recall)d(Recall) \tag{9}$$

The computation of Mean Average Precision (mAP) involves two critical steps: (a) determining AP for each class, and (b) calculating their mean value. This process is mathematically defined as:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{10}$$

i represents the class index, and N denotes the number of categories in the training dataset. mAP0.5 refers to the mean average precision across all classes when the IoU threshold is set to 0.5. mAP0.5:0.95 represents the mean average precision calculated over IoU thresholds ranging from 0.5 to 0.95.

D.Comparative Experiments

To verify whether the proposed detection head structure can achieve the best experimental results under the corresponding conditions, we performed a comparative analysis of the detection head based on YOLOv8s using the VisDrone2019 dataset. Under the influence of the BiFPN concept in the neck network, we tested different detection head structures to evaluate their effects on the model.

First, we applied the concept of BiFPN to improve the neck network. Then, we tested the model using different detection head structures. In our experiments, we defined four different models, the first being Baseline Model 1 (YOLOv8s), Model 2 (after improving the neck network, adding a tiny detection head), Model 3 (after improving the neck network, adding a tiny detection head and removing the large detection head), and Model 4 (after improving the neck network, using the default detection head structure). The experimental results show that after adding a tiny detection head and removing the large detection head, the model achieved the highest average precision, and the number of model parameters significantly decreased, as shown in Table III. Model 2, which added a tiny detection head, also achieved good results but had a higher parameter count than Model 3. Model 4, which used the original detection head structure after applying BiFPN to the neck network, performed much worse than Model 3. Based on these findings, we chose to add a tiny detection head and remove the large detection head. Although this choice increases the computational cost of the model, it maximizes the accuracy of small target detection.

TABLE III
COMPARATIVE EXPERIMENT OF DETECTION HEADS

Models\Metrics	mAP0.5/%	mAP0.5:0.95/%	Parameters/10^6
Baseline	39.1	23.4	11.1
Model2	43.9	26.8	10.6
Model3	43.9	26.9	7.47
Model4	38.9	23.3	11.1

To demonstrate the advantages of WIoU v3, we performed comparative experiments on YOLOv8s by applying WIoU v3 along with several mainstream loss functions. The results of the experiment are summarized in Table IV. The experimental results show that when WIoU v3 is used, mAP0.5 and Recall of the model increase by 0.2% and 0.5%, compared with CIoU. Although some effects were slightly improved when using other loss functions, the model’s overall performance was best when using WIoU v3, proving the effectiveness of the introduction of WIoU v3.

TABLE IV
COMPARATIVE EXPERIMENT OF LOSS FUNCTIONS

Loss\Metrics	Precision/%	Recall/%	mAP0.5/%
CIoU	50.1	38.4	39.1
DIoU	50.2	38.5	39.2
GIoU	49.6	38.6	38.9
EIoU	49.3	38.2	38.5
WIoU v3	49.5	38.9	39.3

To demonstrate the superiority of the BDS-YOLOv8, two comparative experiments were performed using the VisDrone2019 dataset, where the proposed model was evaluated against both YOLO series architectures and other mainstream models. Table V lists the accuracy, mean average precision, recall, parameters, and computational cost of the BDS-YOLOv8 compared to YOLO series architectures. YOLOv3-tiny adopts a simplified network structure to reduce computational complexity and parameter count but sacrifices model accuracy. YOLOv5 adopts the Focus module to expand the receptive field, improving the robustness of the network. YOLOv6 employs an updated self-distillation strategy [25], simplifying the SPPF module in YOLOv5. YOLOv8 improves the Anchor-Free design and optimizes multi-scale feature fusion. YOLOv8, YOLOv9 [26], and YOLOv10 [27], as the cutting-edge versions of the YOLO series, have achieved enhanced detection accuracy. Compared to the YOLOv8s, the BDS-YOLOv8 demonstrates significant improvements across all metrics, achieving 5% higher accuracy, 4.1% greater recall, 5.6% improved mAP0.5, and 3.7% better mAP0.5:0.95, while reducing parameter count by 17.5%.

As illustrated in Table VI, the BDS-YOLOv8 dominates other mainstream models in terms of detection effectiveness. CenterNet [28] proposes an anchor-free detection approach to mitigate the imbalance issue introduced by anchor-based methods, particularly for medium and small-sized objects. Faster R-CNN adopts a Region Proposal Network (RPN), remarkably improving the detection speed of the algorithm. QueryDet accelerates inference for feature pyramid-based dense object detectors using a novel query mechanism, Cascaded Sparse Queries (CSQ), which pre-screens regions containing small objects using high-level low-resolution features, improving small target detection performance to some extent. RetinaNet [29] introduces focal loss, which has been extensively applied to UAV aerial object detection tasks. The EDGS-YOLOv8 [30] model, based on YOLOv8n, excels in model lightweight but sacrifices detection accuracy. ATSS [31] focuses on determining positive and negative samples.

TABLE V
COMPARATIVE EXPERIMENT OF YOLO SERIES ARCHITECTURES

Models\Metrics	Precision/%	Recall/%	mAP0.5/%	mAP0.5:0.95/%	Parameters/10^6	GFLOPs
YOLOv3-tiny	38.1	24.5	23.4	13.1	12.1	18.9
YOLOv5s	49.2	38.2	38.4	22.9	9.11	23.8
YOLOv6	39.5	30.1	29.1	16.9	4.23	11.8
YOLOv8s	50.1	38.4	39.1	23.4	11.1	28.5
YOLOv8m	53.4	41.1	42.3	25.9	25.8	78.7
YOLOv9s	51.4	37.6	39.1	23.6	7.17	26.7
YOLOv10s	49.3	38.2	39.0	23.4	8.04	24.5
BDS-YOLOv8	55.1	42.5	44.7	27.1	9.16	37.0

TABLE VI
COMPARATIVE EXPERIMENT OF VARIOUS MAINSTREAM MODELS

Models\Metrics	mAP0.5/%	mAP0.5:0.95/%	Parameters/10^6	GFLOPs
CenterNet [28]	33.7	18.8	—	—
Faster R-CNN [5]	37.2	21.9	41.7	187
QueryDet [9]	38.1	23.7	—	—
RetinaNet [29]	19.1	10.6	—	—
EDGS-YOLOv8 [30]	31.3	17.6	—	7.9
ATSS [31]	36.4	22.3	—	—
BDS-YOLOv8	44.7	27.1	9.16	37.0

TABLE VII
ABLATION EXPERIMENT

BiFPN	Head	C2f_DSConv	WIoU v3	Precision/%	Recall/%	mAP0.5/%	mAP0.5:0.95/%	Parameters/10^6	GFLOPs
✗	✗	✗	✗	50.1	38.4	39.1	23.4	11.1	28.5
✓	✗	✗	✗	49.1	38.1	38.9	23.3	11.1	28.2
✓	✓	✗	✗	54.4	41.8	43.9	26.9	7.47	33.7
✓	✓	✓	✗	54.1	42.4	44.4	27.1	9.16	37.0
✓	✓	✓	✓	55.1	42.5	44.7	27.1	9.16	37.0

E. Ablation Experiment

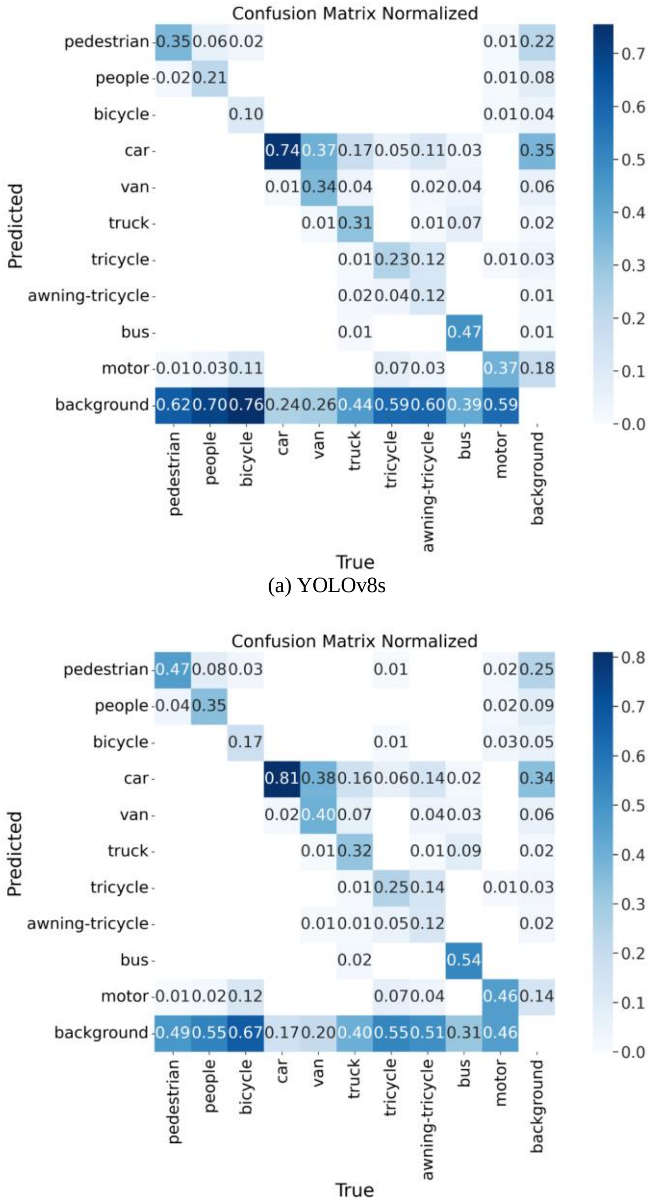
To assess the effectiveness of the proposed enhancement approaches, ablation experiments were carried out on the VisDrone2019 dataset using the baseline model. The corresponding results are illustrated in Table VII. First, after introducing BiFPN in the neck network combined with our proposed tiny detection head structure, mAP0.5 showed a 4.8% improvement, while mAP0.5:0.95 gained 3.5%. Then, replacing the conventional convolutional layer in the Bottleneck with a dynamic snake convolutional layer resulted in 0.5% and 0.2% increases in mAP0.5 and mAP0.5:0.95. Finally, adopting WIoU v3 as the loss function delivered an additional 0.3% boost to mAP0.5. Experimental results demonstrate that BDS-YOLOv8 achieves a 17.5% reduction in parameters compared to YOLOv8s, with concurrent improvements of 5.6% in mAP0.5 and 3.7% in mAP0.5:0.95.

F. Visualization Analysis of Experimental Results

Fig. 8. shows the confusion matrices of YOLOv8s and BDS-YOLOv8. We provide a detailed comparative analysis of the confusion matrices between the baseline YOLOv8s model and the proposed BDS-YOLOv8 model, illustrating the performance enhancement achieved by improving the YOLOv8s model.

The diagonal elements, representing the true positive rate, indicate that BDS-YOLOv8 achieves a higher correct detection rate across all object categories, with a significant improvement exceeding 5% in several challenging categories. This advancement is attributed to the optimized detection head structure and the advanced feature fusion strategy employed in BDS-YOLOv8, which collectively enhance the model’s ability to distinguish target features while effectively suppressing background interference.

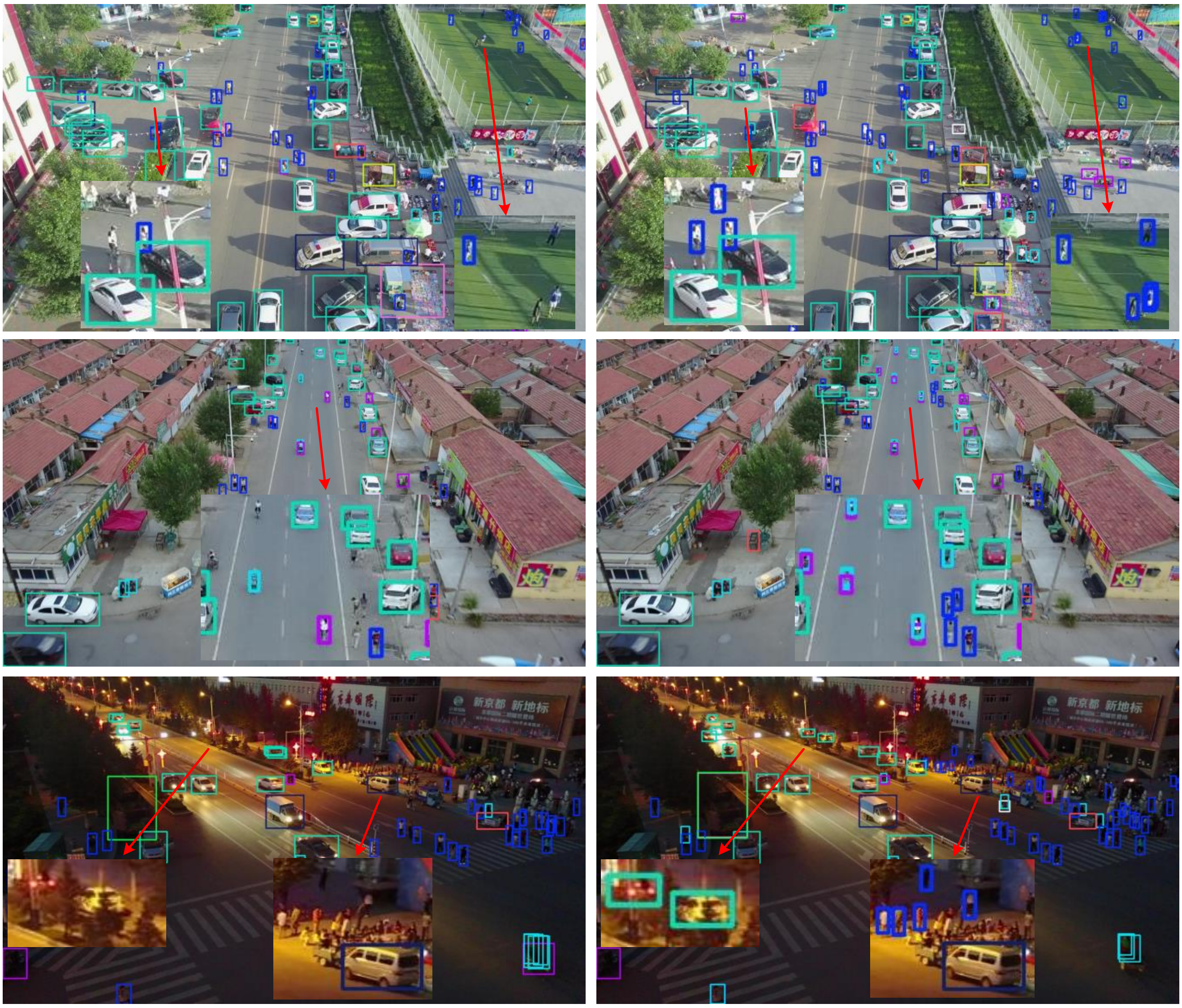
Beyond the diagonal elements, the non-diagonal elements reveal that the improved model significantly reduces inter-class confusion and incorrect background classification, demonstrating its superior robustness in complex scenarios. Reducing background misclassification underscores the model’s improved capability to differentiate foreground objects from noise, a critical factor for practical deployment in real-world applications.



(a) YOLOv8s

(b) BDS-YOLOv8

Fig. 8. Comparison of confusion matrices



(a) YOLOv8s

(b) BDS-YOLOv8

Fig. 9. Comparison of detection performance in different scenarios

A comprehensive comparison of detection results is conducted on UAV aerial images captured from three representative scenarios, as illustrated in Fig. 9. The first row presents a scenario characterized by a high density of small objects under sufficient illumination conditions; the second row represents a sparse distribution of small objects, also under adequate lighting; and the third row depicts a densely populated scene captured at night, reflecting low-light conditions. Across all scenarios, several common object categories—such as bicycles, pedestrians, and vehicles—are present, providing a consistent basis for comparative evaluation.

The experimental results demonstrate that the proposed BDS-YOLOv8 algorithm achieves robust and accurate detection across varying environmental and object density conditions. In both the first and third scenarios, where small objects are densely distributed and often suffer from occlusion, the improved model maintains strong detection capabilities, effectively mitigating the challenges posed by complex spatial overlaps and diverse illumination. Notably, under nighttime conditions (Scenario 3), the model exhibits remarkable resilience to low-light interference, further highlighting its enhanced adaptability and robustness.

In Scenario 2, with a sparse distribution of small targets, compared with the baseline model, BMS-YOLOV8 significantly improves detection accuracy. This indicates that this algorithm can not only effectively handle densely arranged targets but also accurately detect isolated small targets that are easily overlooked. Overall, the proposed improvements are conducive to enhancing the detection stability and generalization. Moreover, the BDS-YOLOv8 has a strong anti-interference ability and broad applicability in unmanned aerial vehicle-based monitoring tasks in the real world.

V.Conclusion

In UAV aerial object detection tasks, challenges such as small object sizes, complex backgrounds, severe occlusions, and varying lighting conditions are prevalent. To address these issues, this paper proposes an improved YOLOv8s algorithm, BDS-YOLOv8. First, a novel detection head structure is designed by adding a tiny-scale detection head that directly utilizes high-resolution feature maps for small target detection while removing the large detection head to reduce parameter overhead. Second, a dynamic snake convolution layer is introduced into the backbone network to improve the model’s ability to capture the features of small objects in complex backgrounds. Additionally, the BiFPN concept is incorporated into the neck network to achieve more efficient multi-scale feature fusion. Finally, WIoU v3 is adopted to substitute the original loss function, optimizing the model’s target positioning accuracy for small targets.

Our experimental results demonstrate that BDS-YOLOv8 significantly improves detection performance compared to the YOLOv8s while reducing the parameters. Specifically, the parameters decreased by 17.5%, mAP0.5 increased by 5.6%, and mAP0.5:0.95 improved by 3.7%. Moreover, compared with mainstream object detection models, BDS-YOLOv8 exhibits notable advantages in both small target detection accuracy and parameters.

Our future research focus is to reduce the computational load of the model further with sustained high accuracy in small target detection and a faster inference speed in the resource-constrained UAV aerial photography scenarios to better meet the practical application requirements.

REFERENCES

[1] D. Wan, M. Zhao, H. Zhou, "Analysis of UAV patrol inspection technology suitable for distribution lines," *Journal of Physics: Conference Series*, vol. 2237, no. 1, p. 012009, 2022.

[2] S. Byun, I. K. Shin, J. Moon, "Road traffic monitoring from UAV images using deep learning networks," *Remote Sensing*, vol. 13, no. 20, p. 4027, 2021.

[3] A. Bouguettaya, H. Zarzour, A. Kechida, "A survey on deep learning-based identification of plant and crop diseases from UAV-based aerial images," *Cluster Computing*, vol. 26, no. 2, pp. 1297-1317, 2023.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, 2009.

[5] S. Ren, K. He, R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks, "*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.

[6] J. Redmon, S. Divvala, R. Girshick, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

[7] W. Liu, D. Anguelov, D. Erhan, "SSD: Single shot multibox detector," *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Oct. 11–14, 2016, Proceedings, Part I, Springer International Publishing*, pp. 21–37, 2016.

[8] Hongyan Li, Baoqing Xu, Ziyang Zhang, and Weifeng Wang, "Small Target Detection Method of Optical Remote Sensing Image Based on Multi-scale Information Fusion," *IAENG International Journal of Computer Science*, vol. 51, no. 6, pp681-687, 2024.

[9] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13668–13677, 2022.

[10] F. Yang, H. Fan, P. Chu, "Clustered object detection in aerial images," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8311–8320, 2019.

[11] C. Xu, W. Lang, R. Xin, "Generative detect for occlusion object based on occlusion generation and feature completing," *Journal of Visual Communication and Image Representation*, vol. 78, p. 103189, 2021.

[12] Lanxue Dang, Gang Liu, Yan-e Hou, and Hongyu Han, "YOLO-FNC: An Improved Method for Small Object Detection in Remote Sensing Images Based on YOLOv7," *IAENG International Journal of Computer Science*, vol. 51, no. 9, pp1281-1290, 2024

[13] Qing Yu, Xinyu Ouyang, Bochao Su, Nannan Zhao, and Hongman You, "Vehicle Detection Algorithm in Complex Scenes Based on Improved YOLOv8," *IAENG International Journal of Computer Science*, vol. 52, no. 4, pp886-893, 2025.

[14] T. Y. Lin, P. Dollár, R. Girshick, "Feature pyramid networks for object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, 2017.

[15] S. Liu, L. Qi, H. Qin, "Path aggregation network for instance segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759–8768, 2018.

[16] Y. Qi, Y. He, X. Qi, "Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6070–6079, 2023.

[17] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 472–480, 2017.

[18] J. Dai, H. Qi, Y. Xiong, "Deformable convolutional networks," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 764–773, 2017.

[19] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10781–10790, 2020.

[20] X. Li, W. Wang, L. Wu, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21002–21012, 2020.

[21] Z. Zheng, P. Wang, W. Liu, "Distance-IoU loss: Faster and better learning for bounding box regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12993–13000, 2020.

[22] Z. Tong, Y. Chen, Z. Xu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," *arXiv preprint arXiv:2301.10051*, 2023.

[23] P. Zhu, L. Wen, D. Du, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, Nov. 2021.

[24] G. Wang, Y. Chen, P. An, "UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios," *Sensors*, vol. 23, no. 16, p. 7190, 2023.

[25] C. Li, L. Li, H. Jiang, "YOLOv6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[26] C. Y. Wang, I. H. Yeh, and H. Y. Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," *European Conference on Computer Vision (ECCV). Cham: Springer Nature Switzerland*, pp. 1–21, 2024.

[27] A. Wang, H. Chen, L. Liu, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107984–108011, 2024.

[28] K. Duan, S. Bai, L. Xie, "Centernet: Keypoint triplets for object detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6569–6578, 2019.

[29] T. Y. Lin, P. Goyal, R. Girshick, "Focal loss for dense object detection," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[30] M. Huang, W. Mi, and Y. Wang, "EDGS-YOLOv8: An improved YOLOv8 lightweight UAV detection model," *Drones*, vol. 8, no. 7, p. 337, 2024.

[31] S. Zhang, C. Chi, Y. Yao, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9759–9768, 2020.