DBYOLOv8: Dual-Branch YOLOv8 Network for Small Object Detection on Drone Image

Yawei Tan[1], Bingxin Xu[2], Jiangsheng Sun[3], Cheng Xu[4], Weiguo Pan[5], Songyin Dai[6], Hongzhe Liu[7]
Beijing Key Laboratory of Information Service Engineering, Beijing Union University, China[1,2,4,5,6,7]
Science and Technology Innovation Research Center, Army Research Academy[3]

***Abstract*—Object detection based on drone platforms is a valuable yet challenging research field. Although general object detection networks based on deep learning have achieved breakthroughs in natural scenes, drone images in urban environments often exhibit characteristics such as a high proportion of small objects, dense distribution, and significant scale variations, posing significant challenges for accurate detection. To address these issues, this paper proposes a dual-branch object detection algorithm based on YOLOv8 improvements. Firstly, an auxiliary branch is constructed by extending the YOLOv8 backbone to aggregate high-level semantic information within the network, enhancing the feature extraction capability. Secondly, a Multi-Branch Feature Enhancement (MBFE) module is designed to enrich the feature representation of small objects and enhance the correlation of local features. Third, Spatial-to-Depth Convolution (SPDConv) is utilized to mitigate the loss of small object information during downsampling, preserving more small object feature information. Finally, a dual-branch feature pyramid is designed for feature fusion to accommodate the dual-branch input. Experimental results on the VisDrone benchmark dataset demonstrate that DBYOLOv8 outperforms state-of-the-art object detection methods. Our proposed DBYOLOv8s achieve mAP@0.5 of 49.3% and mAP@0.5:0.95 of 30.4%, which are 2.8% and 1.5% higher than YOLOv9e, respectively.**

Keywords*—Drone images; dual-branch; small object detection; YOLOv8

I. Introduction

With the development of hardware and artificial intelligence, drones have been gradually applied to intelligent transportation, agricultural monitoring, fire rescue and other fields. In urban traffic monitoring and urban combat missions, UAVs (unmanned aerial vehicle) play an important role by virtue of their advantages such as fast flight speed, high degree of freedom, broad vision and strong adaptability. However, the streets in the city scene have the characteristics of traffic congestion, dense people, and a wide variety of targets. In addition, due to the high-altitude flight of UAV, objects in UAV images are often too small in size and contain limited feature information, which makes it difficult for the network to extract effective features and easy to be lost in the propagation process across the feature layer [1]. In addition, the size of similar objects varies so much that it is difficult for universal object detection methods to effectively locate and identify these objects [2].

Uav object detection is one of the branches of general target detection. According to the processing flow, the target detection algorithm can be divided into two stages and one stage. The two-stage algorithm is characterized by generating a series of regions of interest, and then classifying and regressing these regions. Its advantage is that the two-stage detection algorithm is more detailed, resulting in higher detection accuracy. The disadvantage is that the inference speed is slower than that of single-stage algorithms. The two-stage representative algorithms include Faster R-CNN [3] and Mask R-CNN [4]. The single-stage algorithm extracts the feature information of the target by convolutional neural network, generates the candidate frame, and classifies and locates the target. This detection method consumes less computer resources during inference, and UAV target detection is usually improved based on single-stage algorithm. Single-stage representative algorithms include SSD [5] and YOLO series [6]. YOLOv8 is a commonly used single-phase detection framework, which is often used for various object detection tasks [7]. Its advantage is that the framework is mature, the externally adapted function library is more common, and a variety of inference accuracy improvement tools can be used directly. However, the objects in UAV images often have problems such as small size, complex background environment, and dense area overlap, which limits the ability of the frame to detect small objects. Secondly, the framework is still weak in detecting similar objects at different scales. Therefore, it is necessary to improve the YOLOv8 algorithm to make it suitable for UAV small object detection.

This paper presents a dual-branch small object detection algorithm based on YOLOv8. Firstly, a composite strategy is used to construct auxiliary branches, and the multi-layer semantic information is comprehensively utilized to improve the feature extraction capability of the framework. Second, a multi-branch feature enhancement module is designed, which uses convolution check of different sizes for parallel processing of small object feature information to improve the representation ability of object feature information. In addition, SPDConv can effectively reduce the loss of feature information in the transmission process, which is very effective for small object detection. When it is embedded in the shallow detection branch of the network, the false detection problem can be well improved. Finally, a dual-branch feature pyramid is constructed to deal with the multi-scale change of the target. Experimental results show that the proposed algorithm greatly improves the performance of object detection and can better cope with the requirements of different tasks on model size. Our main improvements and advantages are as follows:

- C2f module is used to construct auxiliary branch, which aggregate multi-high-level semantic information and enhance the feature extraction ability of small objects. SPDConv [13] is used to alleviate the loss of feature information in the downsampling process.

- Multi-branch Feature Enhancement Module (MBFE) is designed to extract small object feature information by using parallel branches of different convolution kernel sizes, which can realize the diversification of small objects feature information expression.
- Dual-branch feature pyramid network (DBFPN) is established for cross-layer connection with YOLOv8 backbone to compensate for information loss caused by feature information transformation.

The structure of this paper is as follows: In Section II, we will briefly introduce our related work to improve the thinking. In Section III, we take a detailed look at the dual-branch YOLOv8 framework. In Section IV, we conduct experiments on a classical drone dataset and provide a detailed analysis of the results. In Section V, we analyze the existing shortcomings and the continued exploration of future work.

II. Related Work

In object detection, the size of the object can be divided into absolute scale and relative scale according to the definition. In the definition of relative scale [8], usually the relative area of all object instances in the same category, that is, the median ratio of the boundary box area to the image area is between 0.08% and 0.58%. However, the way to define small objects based on absolute scale is more widely used, and the MS COCO dataset [9] defines small targets as those with a resolution less than 32 pixels by 32 pixels. The existing methods to solve the small object detection of UAVs can be classified into three categories: (1) By enhancing the feature information of small objects, the network can locate the objects more clearly. (2) Improve the detection accuracy of small objects by improving the ability of network feature extraction. (3) Adopt multi-scale detection strategies to deal with small objects of different sizes.

To improve the ability of network feature extraction, Liang et al. [10]. proposed CBNetV2 network, which uses shallow network to aggregate different high-level semantic information, aiming to enhance the comprehensive application of feature texture features by the model. In addition, they demonstrated experimentally that small objects can be detected more efficiently when shallow features are aggregated only with feature layers higher than this one. This work pioneered the concept of composite backbone networks. Wang et al[11]. proposed Yolov9, the representative of YOLO series, and designed a Programmable Gradient Information (PGI), which uses the characteristics of reversible architecture to retain more input information, thereby reducing the loss of small target feature information. In this framework, the composite backbone network is also constructed. Yan et al. [12]. propose an HCB network that includes a detail extraction backbone (DEB) designed with a smaller acceptance field to better capture details of small objects. This design enhances feature representation without compromising spatial information. However, the above method only uses a single strategy, and because more parameters are often introduced in order to obtain more gradient information, the computational complexity increases and the practical application is limited.

For the enhancement of small object feature information, the detection accuracy of the network can be improved by improving the feature representation of the network for small object and reducing the problem of the loss of small object feature information. Zhang et al. [13]. developed a feature enhancement module specifically for aerial image detection, using the improved FFM module to further capture the context information of small objects, thereby improving the detection accuracy. Raja et al. [14]. designed a step-free convolution to solve the problem of information loss caused by different interpolation calculations for small objects through this lossless downsampling method, thus improving the network's ability to perceive small objects. However, the improved method has been proved by experiments that the network pays too much attention to texture information when applied equally in each feature layer, which leads to the decrease of detection accuracy.

In order to cope with targets of different sizes, Tsung et al. [15]. proposed the concept of feature pyramid. By connecting shallow texture information and high semantic information from top to bottom, the object feature information of each detection layer is enriched. On this basis, Liu et al. [16]. added a new bottom-up path that preserves more detailed information, which is also effective for multi-scale small objects. Tan et al. [17]. used a weighted feature fusion mechanism to give each input feature path a learnable weight, allowing the network to automatically adjust the importance of each path, thus making more efficient use of feature information. By removing invalid nodes and reusing features, the computational cost is reduced, making it suitable for resource-constrained devices. However, the processing method of feature pyramid is only suitable for a single backbone, and for multi-branch networks, the conventional feature pyramid will dilute the target feature information in the concatenation operation.

Although the existing improvement methods have continuously improved the UAV target detection performance, the existing network architecture is still difficult to achieve high precision and multi-task adaptation, especially for the dense area detection problem in the urban scene, and it is urgent to further improve the detection accuracy. Therefore, a variety of improvement methods should be comprehensively used to enhance the detection ability of the detection network for UAV images.

III. DBYOLOv8 Algorithm

A. Overview of YOLOv8

YOLOv8 is an object detection framework based on single-stage deep learning. Compared with the existing version of YOLO series, YOLOv8 can adjust the size of the model by adjusting the scale factor to adapt to different task requirements. Compared to YOLOv10 [18] and YOLOv11, YOLOv8 framework is more mature and has many existing tools to assist reasoning. YOLOv8 network structure is mainly divided into three parts: (1) Backbone network for extracting object feature information. (2) Processing multi-scale features of the feature pyramid pool layer. (3) The detection head of the classification object type information. Fig. 1 shows the schematic diagram of the YOLOv8 algorithm framework. The backbone network extracts the feature information of the object by using the convolution layer of step by step downsampling. The excellent feature extraction ability is the basis of realizing the high-precision object detection algorithm. Path Aggregation-FPN (PAFPN) structure is introduced into the neck structure, and the feature mapping of different scales is combined to enhance the ability of the algorithm to recognize objects of different sizes. The head layer is the main decoupling head structure, which separates the classification and detection head, and becomes the Anchor-Free detection scheme. The YOLOv8 algorithm is widely used in many fields (for example, agricultural inspection, UAV object detection and autonomous driving). However, the performance of the baseline YOLOv8 is not optimal, and there is no targeted design for small objects. In addition, YOLOv8 does not fully combine shallow features and deep features, so that the feature information of small targets is seriously lost in the process of feature transmission. Therefore, the general object detection algorithm framework YOLOv8 is not suitable for small object detection tasks of UAVs. In order to meet the higher task requirements of UAVs for object detection algorithms, it is necessary to improve the existing algorithms by task driving.

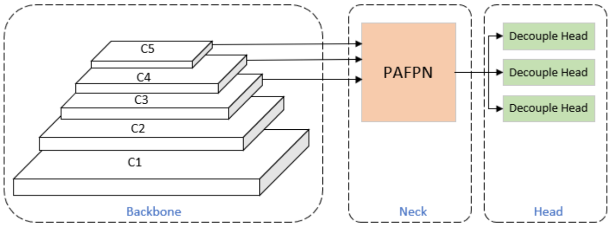


Fig. 1. Simplified diagram of YOLOv8 network structure.

B. Overall Structure of the Optimized DBYOLOv8 Network

In the improved dual-branch YOLOv8 object detection algorithm, taking YOLOv8 as the benchmark model, three aspects of backbone network structure, feature enhancement module and multi-scale feature fusion are optimized and improved. Fig. 2 shows the optimized two-branch YOLOv8 backbone network structure. In the feature extraction stage, this paper designed an auxiliary branch to aggregate the target features of different feature layers. Inspired by the auxiliary branch constructed by CBNet and YOLOV9, the auxiliary branch based on YOLOv8 structure was constructed by using C2F module, which can adjust the size of the model. Based on this method, the constructed DBYOLOv8 model can adapt to the model size requirements of different tasks, and the DBYOLOv8 network model is smaller at the same level of detection accuracy. Feature layer scales the feature map to a fixed size by interpolation, and the small object feature information will be lost in the process of transferring between feature layers. By introducing SPDConv in the shallow layer of the trunk and branches, the problem of small object information loss is alleviated by splitting and reassembling. In addition, EMA [19] module with parallel structure and CBLinear structure are combined to extract small object feature information through different receptive fields of parallel branches. This combination forms MBFE module, does not introduce additional parameters, diversifies the small target feature information, and enhances the generalization ability of the model. The double branch feature pyramid is improved based on BiFPN. By fusing the two-branch feature input with feature weighting, the structure can fuse multi-scale features in the neck network and enhance the model's ability to recognize targets of various sizes and shapes.

C. Auxiliary Branch

In the process of feature extraction, the feature information of small objects is lost or offset to a certain extent with the reduction of the feature map size and the calculation method of interpolation. It constitutes a unique phenomenon, shallow feature is close to the input layer and contains richer texture information, while the deep feature has a larger receptive field and contains more semantic information after multiple convolution. The integrated use of shallow and deep feature can effectively improve the network detection performance [20].

Inspired by CBNetV2, the PGI proposed by YOLOv9 framework builds its auxiliary branch by combining multilevel high-level feature information with shallow feature, hoping to enhance the feature representation capability of the backbone. However, the RepNSCPELAN module is designed to capture

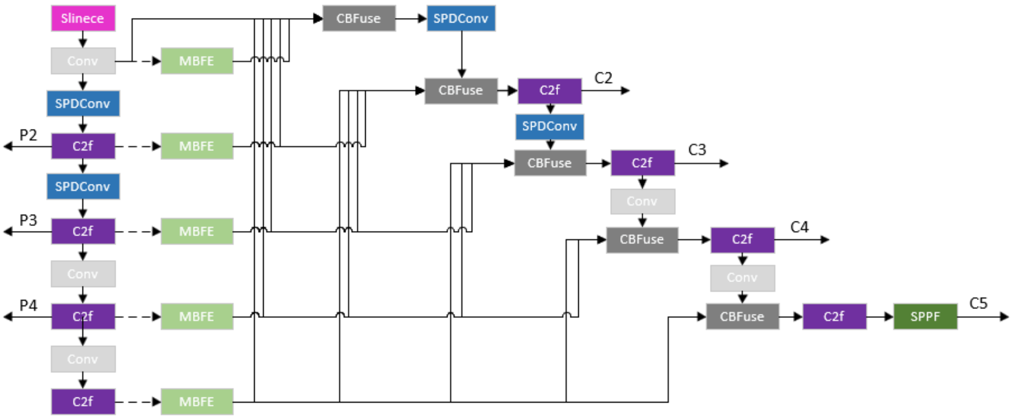


Fig. 2. Our improved DBYOLOv8 feature extraction structure.

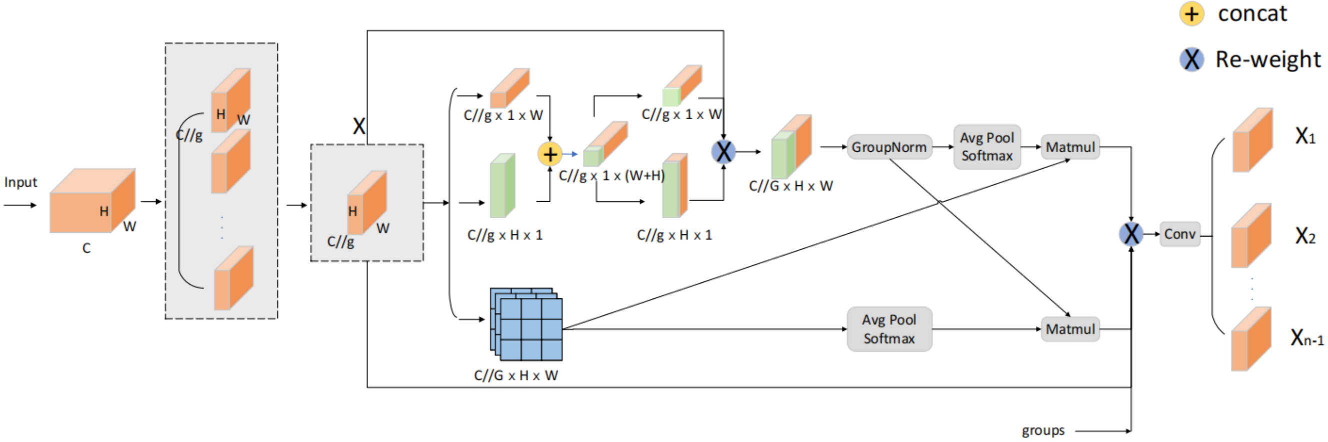


Fig. 3. Multi branch feature enhancement module.

a richer flow of gradient information, which in turn greatly increases the number of model parameters.And it can not adjust the size of its model through scaling factors, making it difficult to adapt to the needs of multiple tasks. In order to improve the feature extraction ability of the framework for small objects without increasing the model size, we built a similar auxiliary branch based on YOLOv8 framework. We use the C2f module to obtain the feature gradient information and the scaling factor to adjust the size of the model to adapt to the task requirements of different platforms.

D. Multi-branch Feature Enhancement Module

The feature information of small objects in UAV images in urban scenes is less, but the background information is complex. Background will seriously affect the extraction of object feature information, which leads to confusion in the transfer process of feature information, thus affecting the performance of the detector.To alleviate this problem, YOLOv9 uses the CBLinear module to process the feature information extracted from the backbone. However, CBLinear module uses the full connection layer to process feature information, which is not friendly for small objects. And it is easy to dilute the feature information of small objects in the process of feature flow, resulting in a certain degree of feature extraction ability loss [21]. Based on the above problems, we design the MBFE module to diversify the small object feature representation.

Channel or spatial attention have been shown to be remarkably effective in producing more recognizable feature representations in various computer vision tasks. In order to diversify the feature information of small objects, we introduced EMA attention mechanism. This mechanism employs multi-branch feature extraction operations, where feature maps are processed in parallel through 1×1 and 3×3 branches. By leveraging different receptive fields, it captures the feature information of small objects and further aggregates the output features of these parallel branches through cross-dimensional interactions to capture pixel-level pairwise relationships.Subsequently, operations such as channel number adjustment and segmentation are performed to obtain a list of multiple output feature maps, which are suitable for subsequent feature aggregation requirements. The designed MBFE module is illustrated in Fig. 3.

E. SPDConv Module

SPDConv proposed by Sunkara et al. is a lossless downsampling method specifically designed for low-resolution images and small objects. Traditional downsampling techniques, such as strided convolutions and pooling layers, often result in the loss of fine-grained information when dealing with low-resolution images or small objects. To address this issue, SPDConv introduces a lossless downsampling method to segment the input image and splicing the input image in the channel dimension so as to retain the input image feature information. After this operation, the convolution layer with stride=0 is used for feature extraction, and the fine-grained details of the image are retained because the size of the feature map is not changed. This approach significantly mitigates the problem of small object loss during the feature extraction phase. We only applied SPDConv during the initial downsampling stages, as deeper features, after multiple convolution operations, have already become highly ambiguous in terms of small object location information. Using SPDConv for downsampling at these deeper stages could negatively impact the detection network [22]. The specific addition location is illustrated in Fig. 1.

F. Dual-branch Feature Pyramid Module

BiFPN removes connection layers that are not intended for fusion and employs a bidirectional weighted strategy to update gradients. Our designed DBFPN is based on BiFPN and is adapted for a dual-branch backbone design. Although the auxiliary branch is constructed based on the yolov8 backbone, after multiple convolutional operations, there is a slight deviation in the mapping of small object feature information and the original image object information. Therefore, incorporating the main branch feature layer information during the construction of the feature pyramid is beneficial for comprehensively utilizing the feature extraction capabilities of both branches. For small object feature information, we have added a P2 detection layer and, considering the need to control the number of

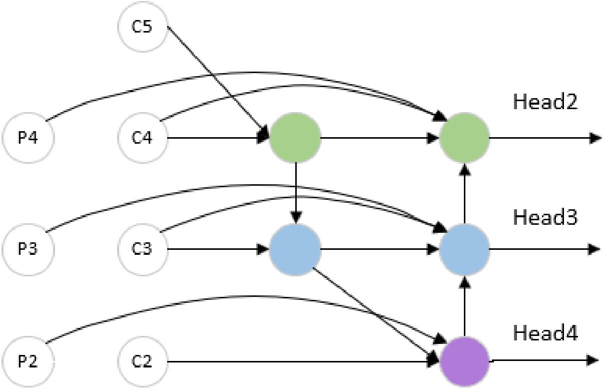


Fig. 4. Dual branch feature pyramid module.

parameters, removed the P5 detection layer [23]. Our designed DBFPN is illustrated in Fig. 4.

IV. Experimental Validation and Analysis

A. Dataset Analysis

In this study, the VisDrone2019 [24] drone object detection dataset, which can represent urban scenes, was used to test the detection performance of DBYOLOv8 on various types of small objects in drone images. The dataset comprises 6,471 images for training and 548 images for validation, with annotations for 10 types of objects, including pedestrians, cars, motorcycles, and others. An analysis of the training set revealed that small objects constitute approximately 60% of the dataset based on their relative scale. Specifically, the dataset categorizes objects as follows: extremely small (es) objects with a length*width in the range [0, 144], relatively small (rs) objects with dimensions in the range [144, 400] pixels, and generally small (gs) objects with sizes in the range [400, 1024] pixels [25]. Given that the dataset was collected using a drone platform, it is particularly well-suited for assessing the performance of the DBYOLOv8 model in detecting small objects from a drone's perspective. Examples of target statistics are shown in Fig. 5.

In order to verify the generalization of our proposed method, a comparative test was also performed on our AI-TOD dataset [26]. AI-TOD dataset is a representative dataset for the detection of tiny objects in aerial images. The dataset contains 28,036 images labeled with a total of 700,621 instances across eight categories (aircraft, Bridges, tanks, ships, swimming pools, vehicles, people, windmills). Compared with other aerial image datasets of the same type, the average size of the object in this dataset is 12.8 pixels, which is much smaller than the object instances in other datasets. Therefore, it is a good way to evaluate the model's perception of small scale objects.

B. Experimental Condition and Assessment Metrics

The experiment was trained and verified on the research group server. The hardware system consists of the following parts: Intel i9 series 13th generation processor I9-13900KF, RTX4090 (24G) graphics card, 64G memory. The software system uses Ubuntu22.04 operating system, and uses Pytorch framework to realize all the algorithms running and improving. For comparison with other algorithms, the input image is set to 640 × 640 pixels and the epoch is set to 200 rounds. The other Settings are the default Settings for the YOLOv8 project provided by the ultralytics team.

To evaluate the algorithm's detection performance on objects in drone images, precision (P), recall (R), mAP@50 and mAP@50:95 were used as evaluation indexes. True positive (TP), false negative (FN), false positive (FP) and true negative (TN) were used as anchor frame positioning quality evaluation. Precision Indicates the percentage of the predicted positive samples that are actually positive. The calculation formula is:

$$precision = \frac{TP}{TP + FP} \tag{1}$$

Recall indicates the proportion of the actual number of positive samples in the total positive samples in which the prediction result is positive. The calculation formula is:

$$precall = \frac{TP}{TP + FN} \tag{2}$$

AP is the Average Precision, which is simply to average the precision value on the PR curve. For the pr curve, we use the integral to calculate. The calculation formula is:

$$AP = \int_0^1 p(r)dr \tag{3}$$

mAP is an evaluation index associated with Intersection over Union (IoU), which averages the detection accuracy of all categories. When loU is set to 0.5, it is usually used as an evaluation index of the detection accuracy of the universal target. mAP@50:95 indicates the mAP with the IoU threshold ranging from 0.5 to 0.95 and the step size of 0.05. Then the average value is obtained. It can also reflect the performance difference of detection algorithms for objects at different scales.

C. Ablation Study

To validate the detection capability of our proposed model for small objects, we constructed auxiliary branches on the basis of the YOLOv8s model, incorporating the SPDconv module, MBFE module, and DBFPN module to build the DBYOLOv8s model. We set the image size to 1280x1280, which is close to the original image size and better reflects the object detection performance of our model on this dataset. The ablation experiment results are shown in Table I. Under the same parameter settings, our method significantly improves the object detection capability for drone images.

1) Effect of auxiliary branches: Small objects occupy a high proportion in drone images and contain limited feature information. To enhance the backbone's ability to extract features from small targets, we constructed an auxiliary branch using the SPDConv module, CBFuse module, and C2f module. By aggregating high-level semantic information from layers not lower than the current one, we enriched the feature information of small targets. Compared to the baseline model, the results for mAP@0.5 and mAP@0.5:0.95 improved by 4.7% and 3.5%, respectively.

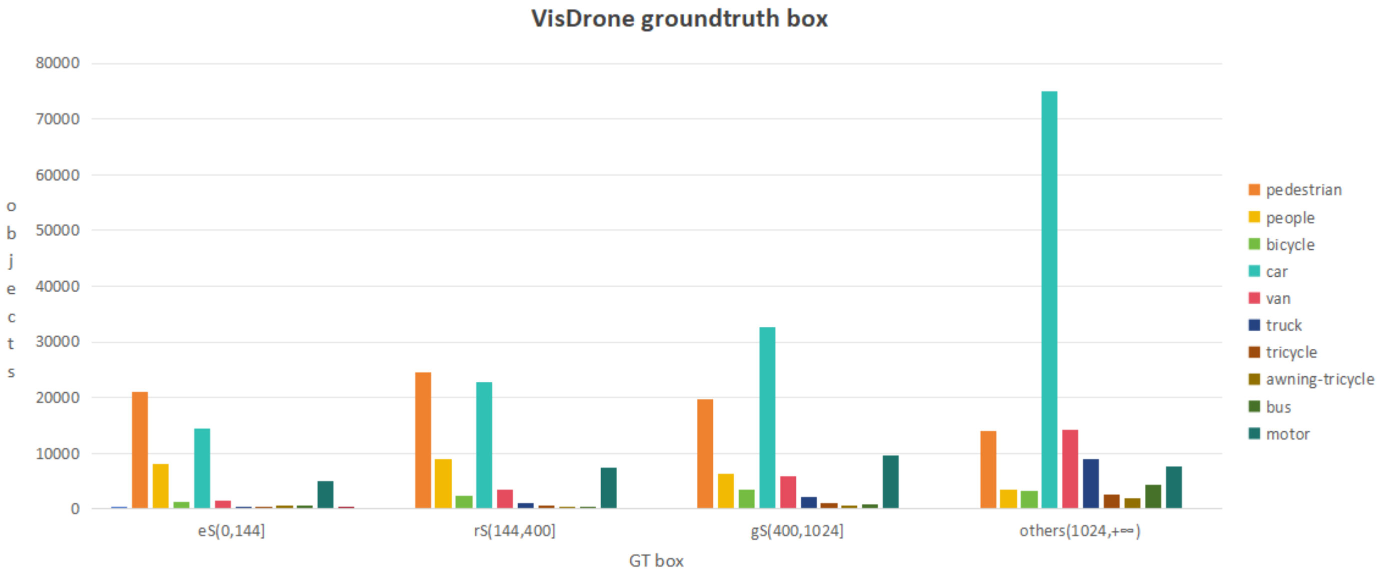


Fig. 5. VisDrone train dataset.

TABLE I. Ablation Study

Baseline	Auxiliary Branch	DBFPN	SPDConv	MBFE	mAP@50(%)	mAP@50:95(%)	Params
✓					56.3	35.4	10.6
✓	✓				61.0	38.9	20.2
✓	✓	✓			61.0	39.0	23.3
✓	✓	✓	✓		61.7	39.5	24.0
✓	✓	✓	✓	✓	62.1	39.9	24.0

2) Effect of DBFPN module: We understand that the feature information after multiple convolutions differs from the original feature information. Moreover, the flow of feature information across layers can result in some information loss. Therefore, our proposed DBFPN integrates dual-branch feature information through skip connections, which improves mAP@0.5:0.95 by 0.1%.

3) Effect of SPDConv module: Small objects may experience varying degrees of feature information loss during the downsampling process due to differences in interpolation methods. As mentioned above, SPDConv can alleviate the issue of feature information loss caused by downsampling in low-resolution images. However, if SPDConv is uniformly applied to replace every downsampling step, it can negatively impact detection performance. This is because, in deeper layers of the network, small objects have less feature information, and the feature information of larger objects is diluted by SPDConv, leading to missed detections. We conducted three sets of experiments: one with SPDConv added to all layers, one with SPDConv added only in the shallow layers, and one with SPDConv added only in the deep layers. The detector achieved the best performance when SPDConv was added only in the shallow layers, as verified by the experiments shown in Table II.

4) Effect of MBFE module: The feature information of small objects is processed through parallel branches, allowing the extraction of target information using convolution kernels of different sizes. This method of enhancing small object features effectively diversifies the representation of small object feature information, thereby enhancing the network's feature extraction capabilities. Compared to the CBlinear module that solely employs fully connected layers, this approach improves mAP@0.5 and mAP@0.5:0.95 by 0.4% without introducing additional parameters.

TABLE II. Comparison Result of Different SPDConv Addition Positions on Thevisdrone2019 Validation Datasets. The Best Result is Highlighted in Bold

Method	mAP@50(%)	mAP@50:95(%)
P1− >P3	**61.1**	**39.2**
P3− >P5	60.1	38.4
P1− >P5	60.9	39.0

D. Comparison with State-of-the-Arts

Due to the varying size constraints for tasks across different platforms, we designed two DBYOLOv8 models of different sizes based on the YOLOv8s and L models. The scaling factors for the DBYOLOv8s model are [0.35, 0.50], while those for the DBYOLOv8L model are [1.00, 1.00]. We compared DBYOLOv8 with other widely used object detection algorithms (primarily the s and l models of various object detection

TABLE III. Comparison Results of Different Object Detectors on Thevisdrone2019 Validation Datasets. The Best Result is Highlighted in Bold

Method	Inputsize	mAP@50(%)	mAP@50:95(%)	Params(M)	FLOPs(G)
RetinaNet[27]	1333*800	39.3	21.8	-	524.95
Faster-RCNN	1333*800	43.6	24.8	-	322.25
YOLOv5-s[28]	640*640	32.2	17.5	7.2	16.5
TPHYOLOv5-s[29]	640*640	37.4	21.7	-	-
YOLOv8-s	640*640	37.3	22.1	11.1	28.5
Drone-YOLO[30]	640*640	44.3	-	10.9	-
yolov10s	640*640	41.2	24.8	8.0	24.5
YOLOv11s	640*640	41.6	25.2	9.4	21.3
HIC-YOLO[31]	640*640	44.3	26.0	-	-
YOLOv5-l	640*640	42.9	26.3	46.5	109.1
YOLOv8l	640*640	43.7	26.7	43.6	165.4
TPHYOLOv5-l	640*640	41.8	24.0	-	-
YOLOv8-x	640*640	44.3	27.2	68.2	258.5
YOLOv9e	640*640	46.5	28.9	57.3	189.0
DBYOLOv8-s	640*640	49.3	30.4	24.0	119.8
DBYOLOv8-l	640*640	**54.4**	**34.3**	175.7	877.0

frameworks). The results, as shown in Table III, indicate that DBYOLOv8 achieved the best and second-best results in terms of mAP. Compared to YOLOv8l and YOLOv9e, DBYOLOv8s achieved higher mAP5@50:95 values by 3.7% and 1.5%, respectively, but with significantly fewer parameters. DBYOLOv8 demonstrated superior performance in small object detection compared to other methods, and the experimental results confirm the competitive advantage and effectiveness of this approach.

To verify the effectiveness of the proposed method in identifying complex backgrounds, significant scale differences, and densely packed small objects, we provide visual examples of DBYOLOv8s and YOLOv8l in Fig. 6. In the first line of the image, the vehicles on the far side of the street are extremely small in size. Carefully examining the red box, it is clear that YOLOv8l cannot fully recognize these extremely small objects, while our proposed method also has certain detection accuracy for extremely small objects. The second line of images taken by the drone from a low altitude Angle shows that the green box shows that YOLOv8l missed the target, and the blue box shows that the method incorrectly identified the person as a motorcycle. In contrast, our proposed approach is not affected by these factors. The observations show that our proposed method shows significant advantages over other methods in processing images of this nature.

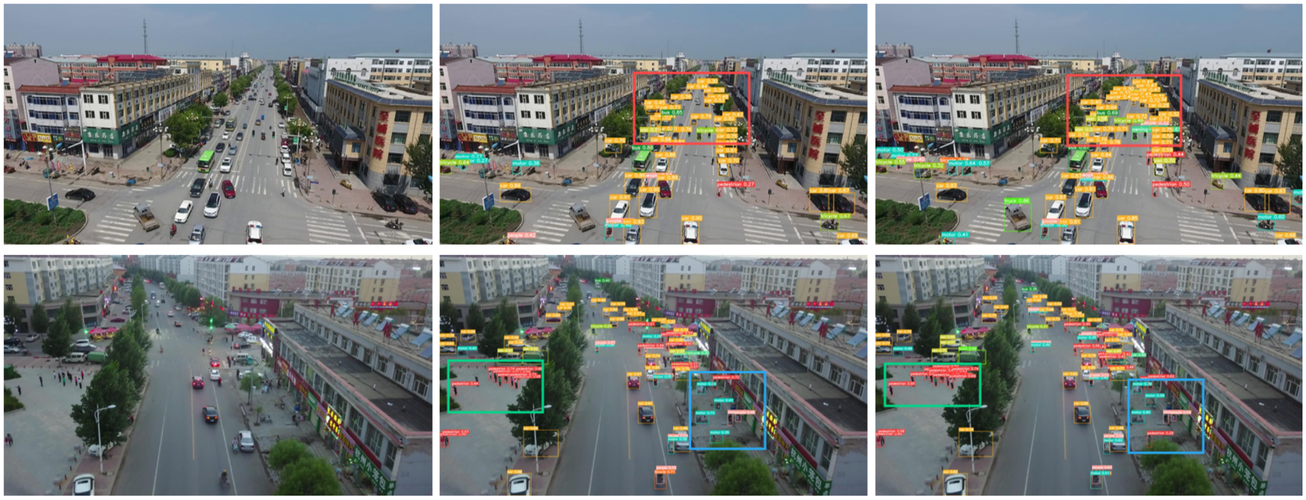
In order to verify the detection ability of the method for small targets and its generalization on other datasets, we conducted training and inference experiments on AI-TOD datasets using the same parameters. Compared with the mainstream YOLO improved algorithm and DERT improved algorithm, our proposed DBYOLOv8s has undoubtedly obtained the best detection results. Experimental results are shown in Table IV. Compared with YOLOv8l, the proposed algorithm at mAP@50:95 improves by 1.2%. Compared with other algorithms, our method also has obvious advantages. Compared with the VisDrone dataset, the AI-TOD dataset has more small object instances, which indicates that our method will improve the detection accuracy if there is more sufficient data support. These results across different datasets underscore the robustness and effectiveness of the proposed method.

TABLE IV. Comparison Results of Different Object Detectors on AI-TOD Validation Datasets. The Best Result is Highlighted in Bold

Method	mAP@50(%)	mAP@50:95(%)
YOLOv6[32]	42.2	18.4
YOLOv7[33]	49.5	19.7
YOLOv8l	48.4	22.0
RT-DETR	48.9	22.7
YOLOv10b	46.7	21.6
YOLOv9c	45.8	20.3
DBYOLOv8-s	**55.2**	**23.2**

V. Conclusion

In order to meet the requirements of existing algorithm frameworks for UAV small object detection, we propose a dual-branch YOLOv8 small object detection algorithm. Firstly, we construct auxiliary branches with compound strategy, combine shallow feature information and higher level semantic information, and increase the feature extraction capability of detection network for small objects. Second, in order to enhance the feature representation of small objects, a multi-branch feature enhancement module is designed to extract the feature information of small objects in parallel through features of different convolution kernel sizes. This module can effectively diversify the representation of small object feature information and counter the problem of the loss of feature information in the process of transmission. Third,



(a) (b) (c)

Fig. 6. Comparison of testing results. (a) Original image. (b) YOLOv8l detection results. (c) Our DBYOLOv8s detection results.

we replace the original subsampling with SPDConv in the shallow layer of the network, and maximize the retention of object feature information through recombination and splicing operations, reducing the missing problem caused by the loss of small and medium-sized object feature information during the subsampling process. Secondly, in order to deal with the contact deviation between the feature information and the original image information caused by multiple convolution, we construct a dual-branch feature pyramid to comprehensively use the double-branch feature information to solve the problem of object scale change in the UAV image. Finally, in addition to using the VisDrone dataset, we also used the AI-TOD dataset to evaluate our proposed approach. The effectiveness of our proposed method is verified by experiments. Compared with the basic YOLOv8s, the DBYOLOv8s algorithm proposed in this paper has increased mAP@50 by 12% and mAP@50:95 by 8.3% on the VisDrone dataset, demonstrating excellent performance compared with other object detection algorithms. On AI-TOD dataset, experimental results validate the generalization of our proposed algorithm, and further prove that our proposed algorithm has higher detection accuracy for small objects if there is more sufficient data support. In addition, the DBYOLOv8l built by us based on YOLOv8l has higher detection accuracy, but the model is larger, which is suitable for tasks with higher detection accuracy supported by high-performance computers. Combined with the existing algorithm foundation and research direction, our future research will focus on the following aspects to tackle difficulties: 1. Explore lightweight technology, reduce model parameters by replacing lightweight backbone or model pruning technology, so that the algorithm can be deployed on embedded devices with low power consumption in the future. 2. Research on small object loss function positioning technology, so that the model can improve the positioning accuracy of dense small objects under complex background. 3. Explore the feature description of different architectures for small objects, and combine the dual-branch idea with CNN architecture and Transformer architecture to further improve the detection accuracy of small objects.

References

[1] Jiang, Huiwei and Peng, Min and Zhong, Yuanjun and Xie, Haofeng and Hao, Zemin and Lin, Jingming and Ma, Xiaoli and Hu, Xiangyun, ”A survey on deep learning-based change detection from high-resolution remote sensing images,” *Remote Sensing*, val.14(7), p.1552, 2022.

[2] P. Mittal, R. Singh, and A. Sharma, ”Deep learning-based object detection in low-altitude UAV datasets: A survey,” *Image and Vision Computing*, vol. 104, p. 104046, 2020.

[3] S. Ren, K. He, R. Girshick, and J. Sun, ”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1137–1149, Jun. 2017

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ”Mask R-CNN,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980-2988

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, ”Ssd: Single shot multibox detector,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer*, 2016, pp. 21–37.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ”You Only Look Once: Unified, Real-Time Object Detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, , pp. 779–788

[7] G. Jocher, A. Chaurasia, and J. Qiu. (2023). *Ultralytics YOLO (Version 8.0.0).* [Online]. Available: https://github.com/ultralytics/ultralytics

[8] Chen, Chenyi, Liu, Ming-Yu, Tuzel, Oncel, and Xiao, Jianxiong. ”R-CNN for Small Object Detection”, in *Computer Vision – ACCV 2016*, pages 214–230

[9] Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C. Lawrence. “Microsoft COCO: Common Objects in Context”. In *Computer Vision – ECCV 2014, Lecture Notes in Computer Science*, pages 740–755, 2014.

[10] T. Liang, X. Chu, Y. Liu, Y. Wang, Z. Tang, W.-T. Chu, J. Chen, and H. Ling, ”CBNetV2: A Composite Backbone Network Architecture for Object Detection,” *Cornell University - arXiv*, Jul. 2021.

[11] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, ”Yolov9: Learning what you want to learn using programmable gradient information,” *arXiv preprint arXiv:2402.13616*, 2024.

[12] Z. Yan, H. Zheng, and Y. Li, ”Detail injection with heterogeneous composite backbone network for object detection,” *Multimedia Tools and Applications*, vol. 81, no. 8, pp. 11621–11637, 2022.

[13] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, ”FFCA-YOLO for small object detection in remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[14] R. Sunkara and T. Luo, ”No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects,” in *Joint European conference on machine learning and knowledge discovery in databases, Springer*, 2022, pp. 443–459.

[15] Lin, Tsung-Yi, Dollar, Piotr, Girshick, Ross, He, Kaiming, Hariharan, Bharath, and Belongie, Serge. ”Feature Pyramid Networks for Object Detection”. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] Liu, Shu, Qi, Lu, Qin, Haifang, Shi, Jianping, and Jia, Jiaya. “Path Aggregation Network for Instance Segmentation”. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[17] M. V. Reddy, K. A. Reddy, M. S. S. Goud, G. Hemanth, and K. Lohith, ”Efficient Det: Scalable and Efficient Object Detection,” *NeuroQuantology*, vol. 20, no. 19, p. 5559, 2022.

[18] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding, ”Yolov10: Real-time endto-end object detection,” *arXiv*,2405.14458, 2024. 1, 3.

[19] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, ”Efficient multi-scale attention module with cross-spatial learning,” in *International Conference on Acoustics, Speech and Signal Processing* , 2023, pp. 1–5.

[20] Yangyang Li, Qin Huang, Xuan Pei, Yanqiao Chen, Licheng Jiao, and Ronghua Shang, “Cross-layer attention network for small object detection in remote sensing imagery”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pages 2148–2161, 2020.

[21] Jiangfan Zhang, Yan Zhang, Zhiguang Shi, Yu Zhang, and Ruobin Gao, ”Unmanned Aerial Vehicle Object Detection Based on Information-Preserving and Fine-Grained Feature Aggregation”, *Remote Sensing*, vol. 16, no. 14, 2024.

[22] Rui Zhong, Ende Peng, Ziqiang Li, Qing Ai, Tao Han, and Yong Tang, ”SPD-YOLOv8: an small-size object detection model of UAV imagery in complex scene”, *The Journal of Supercomputing, Springer*, 2024, pages 1–21.

[23] Lingjie Jiang, Baoxi Yuan, Jiawei Du, Boyu Chen, Hanfei Xie, Juan Tian, and Ziqi Yuan, ”MFFSODNet: Multi-Scale Feature Fusion Small Object Detection Network for UAV Aerial Images”, *IEEE Transactions on Instrumentation and Measurement*, 2024.

[24] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, and Qinghua Hu et al. “VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results”, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 213-226, 2019.

[25] Yuan, Xiang, Cheng, Gong, Yan, Kebing, Zeng, Qinghua, and Han, Junwei. ”Small Object Detection via Coarse-to-fine Proposal Generation and Imitation Learning”. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6294–6304, 2023.

[26] Wang, Jinwang, Yang, Wen, Guo, Haowen, Zhang, Ruixiang, and Xia, Gui-Song. ”Tiny Object Detection in Aerial Images”. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3791–3798, 2021.

[27] Lin, Tsung-Yi, Goyal, Priya, Girshick, Ross, He, Kaiming, and Dollár, Piotr. ”Focal Loss for Dense Object Detection”. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[28]] G. Jocher. (2020). *YOLOv5 By Ultralytics.* [Online]. Available: https://github.com/ultralytics/yolov5

[29] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao, ”TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios”, in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2021.

[30] Z. Zhang, ”Drone-YOLO: an efficient neural network method for target detection in drone images,” *Drones*, vol. 7, no. 8, p. 526, 2023.

[31] Tang, Shiyi, Zhang, Shu, and Fang, Yini. ”HIC-YOLOv5: Improved YOLOv5 For Small Object Detection”. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6614–6619, 2024.

[32] Chuyi Li, Lulu Li, Yifei Geng, Hongliang Jiang, Meng Cheng, Bo Zhang, Zaidan Ke, Xiaoming Xu, Xiangxiang Chu, ”YOLOv6 v3.0: A Full-Scale Reloading,” *arXiv*, preprint arXiv:2301.05586

[33] Wang, Chien-Yao, Bochkovskiy, Alexey, and Liao, Hong-Yuan Mark. ”YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors”. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2023.