IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10642459

YOLOX-DRONE: AN IMPROVED OBJECT DETECTION METHOD FOR UAV IMAGES

Yuqing Zhang, Huanxin Zou, Shitian He, Xu Cao, Meilin Li, Shuo Liu, Liyuan Pan*

College of Electronic Science and Technology,
National University of Defense Technology, Changsha 410073, China

ABSTRACT

Unmanned aerial vehicles (UAV) are widely used for their small size and flexibility. However, the large number of small objects and the significant difference in object size in UAV images bring great challenges to the detection task. Therefore, we propose an object detection method for UAV images with four improvements on the strong baseline model YOLOX-S, which is robust to detect small objects and multi-scale objects. Firstly, we introduce a high-resolution feature map to retain rich detailed information about small objects. Secondly, we propose new up-sampling and down-sampling modules to reduce the feature information loss during the sampling process. Thirdly, we present the triple-scale feature fusion module (TSFFM) to fuse more abundant multi-scale features in the neck's bottom-up feature fusion process. Finally, the parrell dilated convolution attention module (PDCAM) is proposed to learn the multi-receptive field features. Experiment results on the VisDrone-VID2019 dataset validate the effectiveness and superiority of the proposed method.

***Index Terms*—** UAV images, object detection, small object detection, multi-receptive field, context information

1. INTRODUCTION

In recent years, the research of object detection methods in general scenes has made fantastic progress[1, 2, 3]. However, these methods cannot get satisfactory results when directly applied to the UAV scenes.

Compared to object detection in general scenes, object detection in UAV scenes faces greater challenges. Firstly, many objects in UAV images are quite small. Very little feature information is available for object detection. Secondly, the altitude UAVs fly is variable, which makes the size of objects change significantly. To solve the detection difficulties mentioned above, Xi et al. [4] propose a global-local context information collector to extract global and local context information, which can effectively enhance the feature representation of small objects. Li et al. [5] introduce Bi-PAN-FPN to improve the neck part of YOLOV8-S. By fully considering and reusing multi-scale features, a better feature fusion process is realized without increasing much parameter cost. Chalavadi et al. [6] use parallel dilated convolution to learn context information at multiple fields of view. In order to alleviate the sensitivity of IOU to small objects, Yang et al. [7] propose to replace the original IOU-NMS with NWD-NMS in post-processing. A spatial pyramid network, Dp-SPPF, is introduced in [8] to utilize concatenated small-sized max-pooling layers and depth-wise separable convolutions to extract feature information across different scales more effectively.

The above methods fully consider the importance of context information and multi-scale features for object detection in UAV scenes. However, significant feature information attenuation during the sampling process hinders the further improvement of the above methods. Therefore, we propose a new object detection method that comprehensively considers the abovementioned aspects. We choose the YOLOX-S as the robust baseline model. Firstly, a high-resolution feature map is introduced to retain more spatial information of small objects. Secondly, a new up-sampling module called parallel upsample feature fusion (PUSFF) and down-sampling module called parallel downsample feature fusion (PDSFF) are proposed. The improved sampling modules use parallel sample branches to obtain higher quality sampling feature map, effectively reducing feature information attenuation during the sampling process. Thirdly, the TSFFM is proposed to strengthen the multi-scale feature fusion capability of the neck's bottom-up feature fusion network. Finally, referring to Bi-FPN [9], we introduce an extra skip branch between the input node and the output node of PAFPN. The difference is that we add the proposed PDCAM module in the branch to learn multi-receptive field feature information, which helps the network better detect multi-scale objects in UAV images.

2. METHOD

The overview of our method is shown in Fig. 1, which includes a CSPDarknet backbone, an improved Path Aggregation Network with Feature Pyramid Network(PAFPN), and four decoupling detection head. Considering that the shallow feature map contains rich spatial information of small objects, we introduce a high-resolution feature map as the detection layer tailored for small objects. Additionally, we propose some modules to fuse more abundant multi-scale feature

*Corresponding author: Huanxin Zou

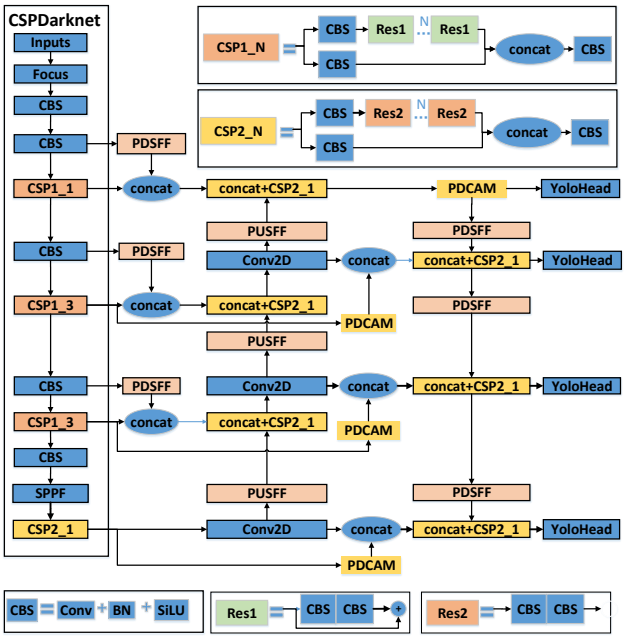


Fig. 1. An overview of our YOLOX-DRONE.

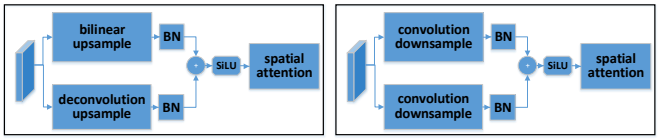
and reduce the attenuation of feature information during the sampling process. Below are the details of each module.

2.1. PUSFF and PDSFF

Most existing feature fusion networks use a single linear interpolation upsampling branch or convolution downsampling branch to adjust the scale of the feature map. There is severe attenuation of feature information during the sampling process [10]. Therefore, we propose the PUSFF and PDSFF, the structures are shown in Fig. 2. Specifically, the PUSFF module consists of two parallel branches. Bilinear interpolation upsampling and deconvolution upsampling operations are performed in these two branches to learn richer upsampling features, and then the upsampling results of the two branches are added for feature fusion. Finally, the spatial attention module is used to make the model focus on the feature information of important areas. Similarly, the PDSFF consists of two downsampling branches. The convolution downsampling branch learns the comprehensive features of the local receptive field, while the max-pooling downsampling branch focuses on the salient features of the original feature map. Lastly, the spatial attention module makes the model focus on the feature information of important areas.

2.2. TSFFM

The shallow layer feature maps contain abundant spatial information, while the deep layer feature maps contain ample semantic information [7]. In the general neck's bottom-up feature fusion process, the feature map of each layer only fuses the feature information of the deeper layer and the current layer, which is rich in semantic information but lacking in detailed information. The lack of detaild imformation is detrimental to the detection of small objects. Therefore, we propose the TSFFM module. The structure is shown in Fig. 3. It can be found that the TSFFM concatenates the shallower layer feature map, the current layer feature map, and the deeper feature map as its output, which contains both rich spatial information and abundant semantic information.



(a) PUSFF (b) PDSFF

Fig. 2. The struction of the PUSFF and the PDSFF.

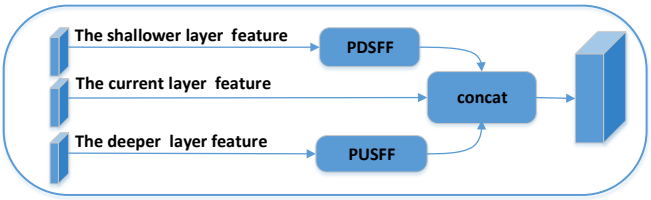


Fig. 3. The struction of the TSFFM.

2.3. PDCAM

Small objects are difficult to detect for limited feature information. However, objects usually exist in a specific environment or coexist with other objects, so our method learns more context information to improve the detection accuracy of small objects. Referring to the idea of BIFPN, we introduce a skip branch between the input node and output node of PAFPN to fuse richer feature information. The difference is that we add the proposed PDCAM in the branch, which help the network learn abundant multi-receptive field information. The structure of PDCAM, as shown in Fig. 4, consists of five parallel branches. The rightmost branch is a convolutional block attention module(CBAM), which enhances the feature information of important regions. The other branches use the 1×1 convolution to reduce the number of channels and then obtain the multi-receptive field features through parallel dilated convolution with different dilate rates. Finally, all features are stacked and fused by a 1×1 convolution operation.

2.4. Loss Function

The loss of our method consists of classification loss, regress loss, and confidence loss. Some difficult samples, such as occluded or small objects, are challenging to detect in UAV

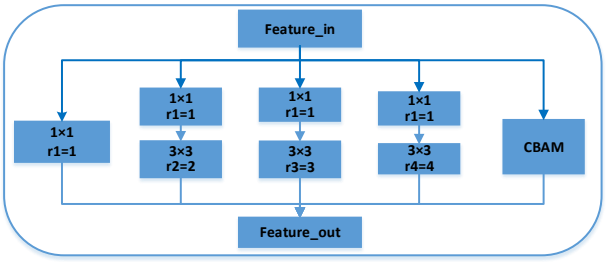


Fig. 4. The struction of the PDCAM.

scenes. Therefore, we choose the focal loss [11] as the confidence loss. The confidence loss function is donated as,

$$L_{obj} = -(\alpha y(1-p)^{\gamma}\log(p) + (1-\alpha)(1-y)p^{\gamma}\log(1-p)), \quad (1)$$

where y indicates the confidence label, p indicates the prediction confidence. α is the loss weight of the positive samples, γ is the focusing parameter, set to 0.75 and 2, respectively.

In the previous methods, the classification score and intersection over union (IOU) score are trained separately. In the inference process, there may be some positions with high classification scores but low IOU scores. Therefore, we choose the quality focal loss (QFL) [12] as the classification loss function, which use continuous labels that combine classification and localization quality. The QFL function is s defined as follows:

$$L_{cls} = -|y-\delta|^{\beta}((1-y)\log(1-\delta) + y\log\delta), \quad (2)$$

where y denotes the continuous label obtained from the IoU of the bounding box and the ground truth. δ indicates the predicted result processed by the sigmoid function. β denotes the hyperparameter of the dynamic scale factor, set to two by default. The expression for L_{reg} is s defined as follows:

$$L_{reg} = 1 - (\frac{|G_t \cap P|}{|G_t \cup P|})^2, \quad (3)$$

where G_t, P denote the ground truth box, the prediction box, respectively. The total loss function L_{total} is denoted as,

$$L_{total} = \lambda_1 L_{reg} + \lambda_2 L_{cls} + \lambda_3 L_{obj}, \quad (4)$$

where λ_1,λ_2 and λ_3 are weights of regression loss, classification loss and confidence loss, set to five,one and one by default, respectively.

3. EXPERIMENTS

3.1. Experimental Settingss

To evaluate the effectiveness of our method, we conduct extensive experiments on the Visdrone 2019-VID dataset[13]. The dataset consists of 10 categories and includes 24,198 images for training and 6635 images for testing. The images are resized to 640×640. For a fair comparison, All methods are run on a single NVIDIA GeForce 3080 GPU and trained for 150 epochs. The optimizer is SGD. Momentum, weight decay, and batch size are set to 0.937, 0.0005, and 4, respectively. The initial learning rate is 0.01, and the cosine annealing learning rate scheduler is used after five epochs of warming up. The data augmentation methods used in the training process include image horizontal flipping, mosaic, and mixup.

Table 1. Results of the comparison methods on Visdrone-VID2019 dataset.The highest performance is **bolded**

Model	mAP	mAP_{50}
YOLOX_S	10.1%	23.2%
YOLOV5_S	11.8%	25.2%
YOLOX_M	12.0%	26.0%
YOLOV6_T	13.20%	27.7%
YOLOV6_S	13.30 %	30.6%
YOLOV7_T	11.20%	25.1%
YOLOX-DRONE(Ours)	**14.30%**	**30.7%**

Table 2. The effectiveness analysis of the proposed modules.The highest performance is **bolded**.

PDCAM	PUSFF+PDSFF	TSFFM	mAP_{50}
-	-	-	27.9%
✓	-	-	28.2%
✓	✓	-	29.4%
✓	✓	✓	**30.7%**

3.2. Comparison to the State-of-the-art Methods

The detection results of the proposed method and all comparison methods on VisDrone-VID2019 are shown in Table 1. It can be found that the mAP_{50} of proposed method has a growth rate of 7.5% compared with the base model YOLOX_S. Compared to other state-of-the-art methods, the proposed method still obtains the most accurate detection results. The visual results are shown in Fig. 5. Other methods fail to detect small targets in the areas marked by the red boxes, but our method can effectively detect these objects, which demonstrates the advantages of the proposed method. It is worth mentioning that compared with YOLOV6, it seems that our method has only a slight improvement in accuracy, but it seems to have a significant effect on the recall rate of small targets. Perhaps focusing on the improvement of classification accuracy is the future improvement direction.

3.3. Ablation Study

To verify the effectiveness of our method, we add the proposed modules to the baseline in turn. The experimental results are shown in Table 2. From the results, we can find that

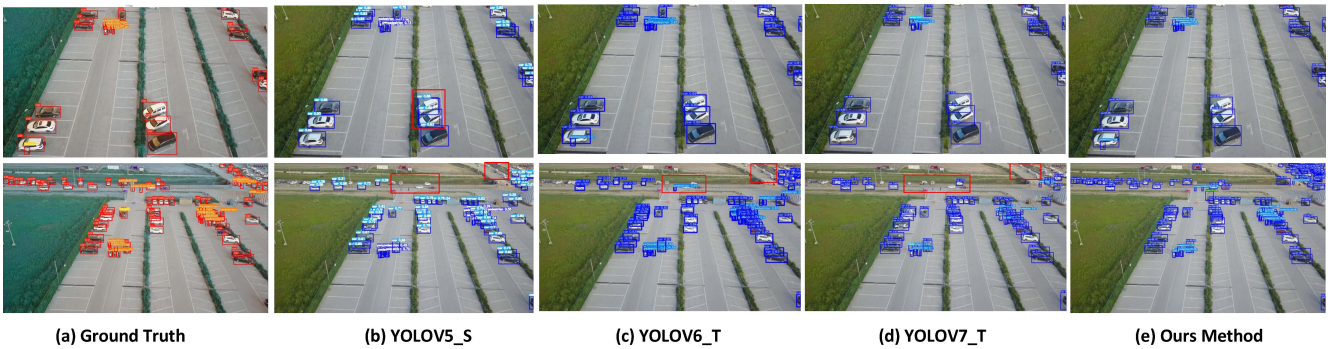


Fig. 5. Visual results of our method and the comparison methods.

the most significant improvement was the addition of TSFFM, with an increase of 1.3% based on the mAP_{50} indicator. After adding the PDCAM, there is a slight improvement of 0.3% based on the mAP_{50} indicator; After adding the improved upsample module PUSFF and downsample module PDSFF, there is an improvement of 1.2% based on the mAP_{50} indicator. It can be found that the detection accuracy increased successively, validating the effectiveness of the proposed modules.

4. CONCLUSION

In this paper, we propose a new object dectection method for UAV images. The framework introduces a high-resolution feature map to retain rich detailed information. The proposed new sampling modules effectively reduce the feature information loss. In addition, by fully learns the context information and multi-scale features of the objects, out method effectively improving the object detection performance in UAV scenes. Experimental results on the VisDrone-VID2019 validate the effectiveness and superiority of the proposed method on UAV image dataset.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant 62071474.

6. REFERENCES

[1] Chuyi Li, Lulu Li, and Jiang et al., "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[2] Chien-Yao Wang, Alexey Bochkovskiy, and Liao et al., "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *IEEE CVPR*, 2023, pp. 7464–7475.

[3] Zheng Ge, Songtao Liu, and Wang et al., "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[4] Yue Xi, Wenjing Jia, and Miao et al., "Fifonet: Fine-grained target focusing network for object detection in uav images," *Remote Sensing*, vol. 14, no. 16, pp. 3919, 2022.

[5] Yiting Li, Qingsong Fan, and Huang et al., "A modified yolov8 detection network for uav aerial image recognition," *Drones*, vol. 7, no. 5, pp. 304, 2023.

[6] Vishnu Chalavadi, Prudviraj Jeripothula, and Datla et al., "msodanet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions," *Pattern Recognition*, vol. 126, pp. 108548, 2022.

[7] Jiale Yang, Hao Yang, and Wang et al., "A modified yolov5 for object detection in uav-captured scenarios," in *2022 IEEE International Conference on Networking, Sensing and Control (ICNSC)*. IEEE, 2022, pp. 1–6.

[8] Yalin Zeng, Tian Zhang, and He et al., "Yolov7-uav: An unmanned aerial vehicle image object detection algorithm based on improved yolov7," *Electronics*, vol. 12, no. 14, pp. 3141, 2023.

[9] Mingxing Tan, Ruoming Pang, and Le et al., "Efficientdet: Scalable and efficient object detection," in *IEEE CVPR*, 2020, pp. 10781–10790.

[10] Yunzuo Zhang, Cunyu Wu, and Zhang et al., "Self-attention guidance and multi-scale feature fusion based uav image object detection," *IEEE GRSL*, 2023.

[11] Tsung-Yi Lin, Priya Goyal, and Girshick et al., "Focal loss for dense object detection," in *IEEE ICCV*, 2017, pp. 2980–2988.

[12] Xiang Li, Wenhai Wang, and Wu et al., "Generalized focal loss: Learning qualified and distributed bounding

boxes for dense object detection," *NIPS*, vol. 33, pp. 21002–21012, 2020.

[13] Pengfei Zhu, Dawei Du, and Wen et al., "Visdrone-vid2019: The vision meets drone object detection in video challenge results," in *IEEE ICCVW*, 2019, pp. 0–0.