Pacific Graphics (2024)
R. Chen, T. Ritschel, and E. Whiting (Editors)

Poster

SPDD-YOLO for small object detection in UAV images

X Xue [1], Y.T Ji[†1], Y Liu[1], H.T Xu[1], Q.D.E.J Ren[1], B Shi[1], N.E Wu[1], M Lu[1], X.F Zhuang[1]

[1]Inner Mongolia University of Technology, China

Abstract

Aerial images captured by drones often suffer from blurriness and low resolution, which is particularly problematic for small targets. In such scenarios, the YOLO object detection algorithm tends to confuse or misidentify targets like bicycles and tricycles due to the complex features and local similarities. To address these issues, this paper proposes a SPDD-YOLO model based on YOLOv8. Firstly, the model enhances its ability to extract local features of small targets by introducing the Spatial-to-Depth Module (SPDM). Secondly, addressing the issue that SPDM reduces the receptive field, leading the model to overly focus on local features, we introduced Deep Separable Dilated Convolution (DSDC), which expands the receptive field while reducing parameters and forms the Deep Dilated Module (DDM) together with SPDM. Experiments on the VisDrone2019 dataset demonstrate that the proposed model improved precision, recall, and mAP50 by 5.8%, 5.7%, and 6.4%, respectively.

CCS Concepts

• ***Computing methodologies*** → ***Object recognition;*** *Object identification;*

1. Introduction

This paper focuses on the common issues of blurriness and low resolution in aerial images captured by drones, which posed challenges for the YOLO object detection algorithm in distinguishing bicycles, tricycles, and similar objects in such scenes. To address this problem, we propose the SPDD-YOLO model based on the YOLOv8 framework. The main contributions of this paper are as follows: • SPDM utilizes spatial information to complement channel information, enhancing local feature extraction capability without losing feature information. • DSDC adjusts the dilation rates at different feature extraction stages to mitigate the gridding effect. • DDM achieves feature fusion through the parallel operation of SPDM and DSDC, enhancing the semantic contrast between targets and backgrounds. Our approach demonstrates superior performance compared to several strong baselines on the VisDrone2019 dataset.

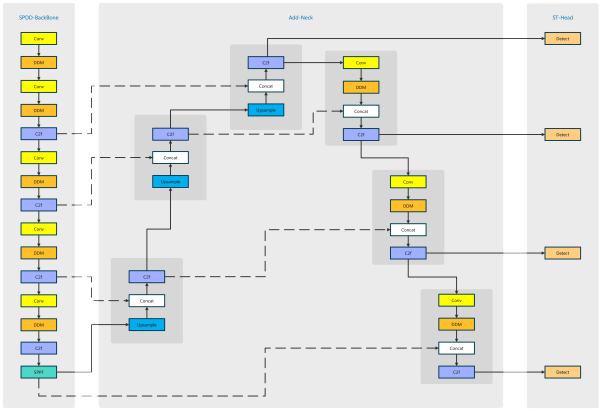

Figure 1: *The network structure of SPDD-YOLO.*

2. SPDD-YOLO

YOLOv8 is a versatile model that can be applied to tasks such as image classification, object detection, and image segmentation. The main changes in YOLOv8 include the use of a new backbone network, anchor-free detection heads, and dropout functions. The primary contribution of this paper is the addition of a lightweight DDM to the YOLOv8 model. The SPDD-YOLO structure in Figure 1 retains the original strengths of the model and enhances its ability to recognize small targets.

SPDD-Backbone: On the backbone of YOLOv8, DDM is added after each Convolutional module. The DDM consists of parallel SPDM and DSDC components and fuses the feature information from both. SPDM processes tensors, resulting in four tensors with the same number of channels as the input tensors but with spatial dimensions halved. These tensors are after that concatenated along the channel dimension. This operation ensures that the size of feature maps remains unchanged in each stage of feature extraction, thereby enhancing the network's capability to extract features from minor objects. DSDC is based on depth-separable convolution [Cho17] and increases the dilation rate, thus expanding the receptive field while reducing the number of parameters and computational load. Due to the insignificant parameter count of SPDM and DSDC, DDM is a lightweight module.

DOI: 10.2312/pg.20241327

Add-Neck: In the feature fusion phase, additional up-sampling and down-sampling steps are introduced on top of the existing framework. This is particularly beneficial for small object detection, improving the precision and accuracy of object positioning. The purpose of up-sampling is to increase the resolution of the feature map, allowing the network to better capture information about small objects or distant details. At each down-sampling stage, DDM is incorporated to extract features of small objects while complementing information other than small targets.
ST-Head: In the head section, the anchor-free head and decoupled head of YOLOv8 has been retained and an additional head for smaller objects is introduced.

3. Experiments

The optimizer used in the experiments is not fixed and the model selects an appropriate optimizer based on the number of iterations. All experiments were conducted for 100 epochs with a batch size of 16 (RT-DETR [LXZ*23] used a batch size of 8), and images were resized to 640x640 pixels. The experiments utilized the VisDrone2019 dataset, collected by the AISKYEYE team at Tianjin University's Machine Learning and Data Mining Laboratory.

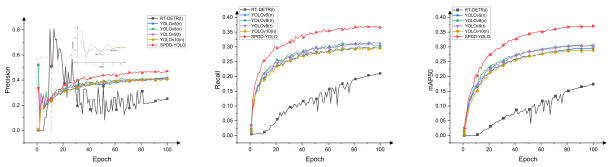

Figure 2: *Visualization compares model performance metrics: Precision, Recall, and mAP50.*

Inspired by [WCY*17], to address the "gridding effect" issue, in the backbone, the dilation rates of DSDC in DDM are set to [1,2,3,1]. We compared the performance of the SPDD-YOLO model with various object detection models [Joc20].

As can be seen from Figure 2, compared to other target recognition algorithms, the Recall and mAP50 of the proposed SPDD-YOLO model showed a significant improvement by the 20th epoch. For precision, it is evident that the YOLOv8 model exhibits significant instability, repeatedly oscillating throughout the initial 8 epochs. However, SPDD-YOLO markedly reduces this oscillation amplitude and resolves this state sooner. Concurrently, by the 20th epoch, SPDD-YOLO achieves significantly higher precision compared to YOLOv8.

Table 1: *Comparison of performance among different models*

Methods	P(%)	R(%)	mAP50(%)	mAP50-95(%)
RT-DETR(l)	26.3	21.1	17.4	9.56
YOLOv5(n)	40.4	29.7	29.5	16.8
YOLOv8(n)	41.1	31.2	30.7	17.6
YOLOv9(t)	41.3	31.3	30.5	17.6
YOLOv10(n)	41	29.2	28.9	16.5
SPDD-YOLO	**46.9**	**36.9**	**37.1**	**21.9**

Table 1 demonstrates the superiority of our algorithm on the VisDrone2019 dataset. Compared to other object detection methods, SPDD-YOLO outperforms in all metrics. Its precision, recall, mAP50, and mAP50-95 are 46.9%, 36.9%, 37.1%, and 21.9%, respectively. As shown in Figure 3, compared to the better-performing YOLOv8, this algorithm additionally identifies distant targets, distinguishes between bicycles or tricycles, and enhances the probability of recognizing desired targets. However, the performance remains inadequate for targets with low contrast, which could be a direction for future research.

It is evident from figures, tables, and comparative experiments that the SPDD-YOLO model significantly enhances the recognition ability of small objects, surpassing all metrics of previous state-of-the-art (SOTA) models.

Figure 3: *Three images represent the original image, the results recognized by YOLOv8, and the results recognized by SPDD-YOLO. We set the confidence level to 0.41.*

4. Conclusion

This paper addresses the problem of low-altitude UAV target recognition and proposes the SPDD-YOLO model. This model is based on YOLOv8 and incorporates the DDM, which focuses on small targets while enlarging the receptive field. Comparative experiments show that the SPDD-YOLO model performs excellently on the VisDrone2019 dataset. Thus SPDD-YOLO is suitable for UAV aerial object detection tasks.

5. Acknowledgments

This study is supported by the National Natural Science Foundation of China (62206138, 62066035), Inner Mongolia Natural Science Foundation (2024MS06009), Education Department Science Research Foundation of Inner Mongolia Autonomous Region (JY20220186, NJZZ23081, RZ2300001743), Science Research Foundation of Inner Mongolia University of Technology (BS2021079, ZZ202118).

References

[Cho17] CHOLLET F.: Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 1251–1258. 1

[Joc20] JOCHER G.: Ultralytics yolov5, 2020. URL: https://github.com/ultralytics/yolov5, doi:10.5281/zenodo.3908559. 2

[LXZ*23] LV W., XU S., ZHAO Y., WANG G., WEI J., CUI C., DU Y., DANG Q., LIU Y.: Detrs beat yolos on real-time object detection, 2023. arXiv:2304.08069. 2

[WCY*17] WANG P., CHEN P., YUAN Y., LIU D., HUANG Z., HOU X., COTTRELL G.: Understanding convolution for semantic segmentation. *arXiv preprint arXiv:1702.08502* (2017). 2