Scholar Publishing Group
International Journal of Multimedia Computing
https://doi.org/10.38007/IJMC.2025.060105
ISSN 2789-7168 Vol. 6, Issue 1: 56-63

SPD-YOLO: Enhanced Small Object Detection for Drone Imagery

GuangJun Lai[1,a], HaiMin Wang[1,b], Huiheng Suo[1,c], TingQi Zhou[1,d], Meng Qin[1,e], Qingyuan Xiao[1,f], Zuteng Chen[1,g], Jian Wu[1,h,*], Yuanhao Pan[2,i], Yingping Bai[3,j], Qinxi Lin [4,k]
[1]*Nanchang Hangkong University, Nanchang, China*
[2]*Ningbo University of Finance & Economics, Ningbo, China*
[3] *NingboTech University, Ningbo, China*
[4]*Zhejiang Tianchen Test and Control Tech Co. , Ltd, Cangnan, China*
[a]*l2647080898@163.com*, [b]*2306602311@qq.com*, [c]*suohuiheng@163.com*, [d]*1169575140@qq.com*, [e]*alphenqin@163.com*, [f]*xiaoqy0918@163.com*, [g]*2293747855@qq.com*, [h]*flywujian@qq.com*, [i]*1090824061@qq.com*, [j]*3459951916@qq.com*, [k]*2205605053@qq.com*
[*]*Corresponding author*

Keywords: small target detection; SPD-Conv; UAV aerial photography; YOLOv8n; WIoUv3

Abstract: Aiming at the key technical challenges of small target detection in UAV aerial photography scenarios, this study proposes an improved scheme SPD-YOLO based on the YOLOv8n architecture. The scheme achieves performance breakthroughs through three core innovative modules: 1) adopting the SPD-Conv module instead of the traditional downsampling operation to maintain the resolution of the feature map through the spatial pyramid decomposition strategy; 2) introducing the P2 high-resolution feature layer to construct an enhanced feature pyramid, which improves the feature extraction capability of tiny targets; 3) adopting the WIoU v3 loss function to optimize the positioning accuracy through the dynamic focusing mechanism. Experiments on the VisDrone2019 test set demonstrate that the complete solution (SPD-Conv + P2 + WIoUv3) achieves an mAP@0.5 of 38.3%, surpassing the baseline YOLOv8n by 5.3 percentage points, with precision and recall reaching 49.1% and 37.2%, respectively. Ablation experiments validate the effectiveness of each module: the introduction of the P2 feature layer alone improves 2.6 percentage points, combined with WIoU v3 improves another 1.2 percentage points, and finally the introduction of the SPD-Conv module improves the overall performance by 5.3 percentage points. This scheme significantly improves the detection performance of small targets in UAV aerial photography scenarios while maintaining real-time detection speed.

1. Introduction

Target detection technology, as a core research direction in the field of computer vision, has demonstrated important application value in many key areas such as security monitoring, intelligent

transportation, industrial quality inspection and automatic driving. In recent years, with the rapid popularization of UAV technology, its unique advantages in urban governance, environmental survey, agricultural production, disaster warning and other scenarios have become increasingly prominent. With its high-altitude overlooking perspective and fast response capability, UAV can efficiently acquire high-definition image data in a wide range of areas, which provides a new way of data acquisition for various applications.

In the development of target detection algorithms, deep learning-based methods have mainly formed two technical routes: the two-stage detection framework represented by R-CNN generates the candidate region first and then carries out classification and regression, which has high accuracy but low computational efficiency; while the single-stage detector represented by the YOLO series adopts an end-to-end detection method, which accomplishes feature extraction and target localization, which has a significant advantage in real-time. It is worth noting that RetinaNet effectively solves the category imbalance problem by introducing Focal Loss [1], while the YOLO series algorithms achieve a better balance between detection accuracy and computational efficiency.

However, the existing algorithms still face many challenges in UAV aerial photography scenarios [2], especially the performance for small target detection needs to be improved. To address this problem, researchers have proposed a variety of innovative solutions. liu et al. improve the utilization of shallow features by optimizing the feature extraction strategy of SSD [3]; Wang's team introduces the STDS structure and combines it with the BiFormer attention mechanism in YOLOv8[4], which ensures the detection performance while controlling the complexity.

2. YOLOv8n algorithm introduction

YOLOv8 as the latest evolution of YOLO series algorithms, is officially released by the ultralytics team in 2023, which represents the latest technology in the field of real-time target detection. The algorithm inherits the excellent architecture of YOLOv5, and realizes the comprehensive improvement of detection performance through several innovative improvements.YOLOv8 adopts the backbone network structure based on the CSP (Cross Stage Partial) idea [4], combined with the improved Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) [5] to perform multi-scale feature fusion, which effectively enhances the model to detect targets of different sizes. It effectively enhances the detection ability of the model for targets of different sizes. The core innovation is the introduction of a new C2f module, which draws on the ELAN structure design concept of YOLOv7 [6], and significantly improves the feature expression capability of the network by optimizing the feature transfer path. In the design of the detection head, YOLOv8 adopts the strategy of separating the classification head and the detection head, and innovatively adopts the anchor-free mechanism to directly predict the edge position of the target, this improvement not only simplifies the model structure, reduces the computational complexity, but also effectively reduces the interference of the label noise. To meet the needs of different application scenarios, YOLOv8 provides multiple versions from lightweight to high performance (including YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x), which achieve faster inference speed while maintaining excellent detection accuracy.

3. Improvements to YOLOv8n

3.1 Introduction of SPD-Conv

In target detection tasks, the main challenges of small target detection are its limited pixel occupancy and insufficient feature information. When small and large targets appear simultaneously, model training is easily dominated by the large target, leading to insufficient learning of the small

target. Traditional convolutional neural networks further lose the detailed features of small targets when performing feature extraction through stepwise convolution and pooling operations, thus affecting the detection performance.

To address this problem, this study introduces the *SPD-Conv* module into the backbone network of YOLOv8n, replacing the original step-spanning convolution and pooling operations.*SPD-Conv* consists of a space-to-depth transformation (SPD) layer and a non-step-spanning convolutional layer. Among them, the SPD layer achieves downsampling by reorganizing the spatial information while retaining the complete channel information, while the non-stepwise convolution layer extracts richer feature representations by using the extended channel dimension without changing the feature map size. This improvement effectively alleviates the problem of information loss of small target features during downsampling, thus enhancing the detection accuracy.

For any feature map X with dimension $S\times S\times C1$, it can be decomposed into several sub-feature maps. The specific realization is as follows:

$$f_{0,0} = X[0:S:scale, 0:S:scale],$$

$$f_{1,0} = X[1:S:scale, 0:S:scale],$$

$$f_{scale-1,0} = X[scale-1:S:scale, 0:S:scale],$$

$$f_{0,1} = X[0:S:scale, 1:S:scale],$$

$$f_{1,1} = X[1:S:scale, 1:S:scale],$$

$$\vdots$$

$$f_{scale-1,1} = X[scale-1:S:scale, 1:S:scale],$$

$$f_{0,scale-1} = X[0:S:scale, scale-1:S:scale],$$

$$f_{scale-1,scale-1} = X[scale-1:S:scale, scale-1:S:scale],$$

For any input feature map X,its sub-map $f_{x,y}$ consists of all elements whose position coordinates are divisible by the scale factor $scale$.

The feature map $X(S,S,C_1)$ is transformed to $X'(S/scale, S/scale, scale^2C_1)$ by a space-to-depth transformation (SPD). a convolutional layer with *stride=1* (filter number $C_2 < scale^2C_1$) is subsequently connected to further obtain $X''(S/scale, S/scale, C_2)$. The use of a non-spanning convolution (*stride=1*) maximizes the preservation of the feature information: if a convolution with *stride>1* is used, it can be straightforward to achieve the $X \rightarrow X''$ size transformation, it leads to asymmetric sampling and non-selective loss of feature information.

3.2 Introducing of the WIoUv3 loss function

Aiming at the characteristics of UAV aerial images with many small targets and serious occlusion, the traditional detection methods often perform poorly.The CIoU loss function adopted by YOLOv8 has some limitations although it improves the detection effect by introducing three constraints, namely, overlap area, centre point distance and aspect ratio. Its calculation formula is:

$$L_{CIoU} = L_{IoU} + \frac{\rho^2(b,b_{gt})}{c^2} + \alpha v \quad (1)$$

$$L_{IoU} = 1 - \text{IoU}, \text{IoU} = \frac{|A\cap B|}{|A\cup B|} \quad (2)$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w_{gt}}{h_{gt}} - \arctan\frac{w_{\text{b}}}{h_{\text{h}}})^2 \quad (3)$$

$$\alpha = \frac{v}{(1-IoU)+v} \tag{4}$$

The CIoU loss function optimises the bounding box regression accuracy in target detection through multi-dimensional constraints, and each parameter in its mathematical expression represents: the intersection and concurrency ratio (IoU) reflects the degree of overlap between the predicted box and the real box, the centroid coordinate describes the spatial location of the bounding box, the Euclidean distance measures the degree of centroid offset, the minimum external rectangle diagonal length is used for the normalisation process, the weight coefficient in the aspect ratio penalty term α regulates the penalty intensity, and similarity ratio v measures the aspect ratio difference.When the predicted frame has the same aspect ratio as the real frame ($w_{gt}/h_{gt} = w_b/h_b$), the similarity ratio v=0 invalidates the penalty term, at this time, CIoU can effectively avoid the optimisation stagnation problem due to the same aspect ratio, and improve the model's detection robustness in complex scenes.

In order to improve the generalisation performance of the target detection model when the prediction frame overlaps with the real frame, this study uses a dynamic non-monotonic focusing mechanism to construct a novel loss function. The distance-attention mechanism established by the distance metric forms a WIoUv1 loss function containing a two-layer attention structure. The innovativeness of this method is reflected in the combination of geometric constraints and the attention mechanism, which effectively alleviates the problem of over-penalisation of overlapping regions in the bounding box regression process of the traditional method, and thus improves the adaptive ability of the model in complex scenes.The specific calculations are described below.

$$L_{WIoUv1} = R_{WIoU}L_{IoU}; R_{WIoU} \in [1,e), L_{IoU} \in [0,1.0) \tag{5}$$

$$R_{WIoU} = \exp\left(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{(W_g^2+H_g^2)*}\right) \tag{6}$$

where, R_{WIoU} denotes the regression loss of the high quality anchor frame, (x, y) and (x_{gt}, y_{gt}) represent the centre coordinates of the predicted and real frames, respectively, and W_g and H_g are the width and height dimensions of the smallest outer rectangle of both.This loss function adopts a bidirectional adjustment strategy: on the one hand, it enhances the IoU loss L_{IoU} for normal quality anchor frames through R_{WIoU}, and on the other hand, it uses L_{IoU} to suppress the value of R_{WIoU} for high quality anchor frames.

WIoUv3 is optimised and improved on the framework of WIoUv1, which effectively suppresses the negative gradient effects caused by low-quality samples by introducing dynamic adjustment parameters.WIoUv3 is calculated as follows.

$$L_{WIoUv3} = rL_{WIoUv1};\ r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{7}$$

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \in [0, +\infty) \tag{8}$$

where, non-monotonic focusing coefficient r, outlier β, and hyperparameters α and δ. The dynamic cross-parallel ratio loss L_{IOU}^* serves as the core regulation index. The loss function effectively solves two key problems through an intelligent gradient assignment strategy: firstly, it suppresses the interference gradient generated by low-quality samples, and secondly, it optimises the focus on ordinary quality anchor frames, which significantly improves the accuracy and model generalisation performance of small target detection in UAV aerial photography scenarios, and in particular, it exhibits superior adaptability in the tasks of complex background and tiny target detection.

3.3 P2 microscale target detection layer

The three detection heads of the original YOLOv8 model receive feature maps from the fusion of the backbone network and the feature pyramid network, which have the sizes of 80×80, 40×40, and 20×20 pixels, corresponding to the input image of 640×640 pixels after 8-fold, 16-fold, and 32-fold downsampling. In the UAV small target detection task, since the target size is usually extremely small (e.g., below 10×10 pixels), such deep downsampling leads to a serious loss of small target feature information. To solve this problem, this paper proposes to add a P2 microscale detection layer on top of the original detection architecture, which fuses the 160×160 feature map (containing richer small target and detail information) obtained from the second layer in the backbone network after 4-fold downsampling with the higher-level features. This 160×160 feature map has a smaller sensory field and stronger detailed feature expression capability, which can effectively retain the small-size target information captured by the UAV. By cross-scale feature fusion of the P2 feature map with the P3-P5 feature maps, a four-level detection system containing 160×160 (P2), 80×80 (P3), 40×40 (P4), and 20×20 (P5) is constructed, which significantly improves the model's detection capability for small targets.

4. Experimental results and analysis

4.1 Experimental environment

The experimental environment of this paper is shown in Table 1, the input image resolution is 640×640, the initial learning rate is 0.01, and the attenuation coefficient is set to 0.0001.

Table 1. Experimental environment and parameters

name	parameters
operating system	Linux
CPU	Intel(R) Xeon(R) Platinum 8255C
GPU	NVIDIA GeForce RTX 3080
display memory	10GB
development environment	Python 3.8
development framework	Pytorch 1.11
CUDA	11.3

4.2 Dataset

In this study, the VisDrone2019 dataset [7-8] is used to carry out training, validation and testing experiments. The dataset was acquired by the Machine Learning and Data Mining Laboratory of Tianjin University via UAV, and contains a total of 8629 static images, which are divided into 6471 training images, 548 validation images and 1610 test images. All the images in the dataset cover 10 types of target objects, including pedestrians, occupants (driving tools or stationary people), cars, vans, buses, trucks, motorcycles, bicycles, awning tricycles, and ordinary tricycles, which not only include small target objects such as small vehicles and pedestrians, but also cover diverse and complex scenes.

4.3 Evaluation indicators

For the performance evaluation indexes of the target detection task, this paper adopts Precision,

Recall and mean average precision (mAP) as the core evaluation system, and the definitions and calculation methods of each index are as follows:

4.3.1 Precision:

$$P = \frac{TP}{TP+FP}$$

Where, TP is true cases (number of positive samples correctly detected) and FP is false positive cases (number of negative samples misdetected). This indicator reflects the reliability of the detection results, the higher the value indicates the lower the false detection rate.

4.3.2 Recall:

$$R = \frac{TP}{TP+FN}$$

FN is the number of false counterexamples (positive samples of missed detections). It measures the completeness of the test, with higher values indicating a lower rate of missed tests.

4.3.3 Average Precision (AP)

$$AP = \int P(R)dR$$

Obtained by integrating Precision-Recall curves, this metric integrates the *P-R* trade-off relationship and reflects the detection accuracy of a single category.

4.3.4 Mean Average Precision (mAP)

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k}$$

where *k* is the total number of categories, obtained by averaging the *APs* of each category, as a criterion for the overall performance of the algorithm.

4.4Ablation Experiments

All experiments in this study were conducted under strictly consistent data parameters and environment configurations. In order to verify the effectiveness of the proposed algorithm for detecting small targets in UAV aerial photography scenes, multi-module improvement ablation experiments are conducted on the YOLOv8n base model based on the VisDrone2019 dataset, including the introduction of the WIoUv3 loss function, the SPD-Conv module, and the P2 microscale target detection layer. The experimental design takes YOLOv8n as the base model and constructs comparative experimental groups by progressively integrating the improved modules: group A adds only the P2 detection layer, group B adds the WIoUv3 loss function on the basis of group A, and group C further integrates the SPD-Conv module on the basis of group B to constitute the complete scheme. This progressive experimental design can effectively separate the contribution of each improved module and ensure a fair assessment of the independent and combined effects of

each module. The experimental results are summarised in Table 2.

Table 2. Ablation experiment results

Arithmetic	P2	WIoU v3	SPD-Conv	P/%	R/%	mAP_{50} /%	mAP_{50-95} /%
YOLOv8n				44.8	33	33	19.1
A	√			46.2	35.1	35.6	20.5
B	√	√		47.6	35.9	36.8	21.4
C	√	√	√	49.1	37.2	38.3	22.1

5. Conclusion

In this paper, a target detection algorithm based on the improved YOLOv8n architecture is proposed to address the key issues of large target scale difference, dense spatial distribution and difficult feature extraction in UAV aerial photography scenes. By introducing the SPD-Conv spatial pyramid decomposition convolution module, the algorithm realizes the progressive expansion of the sensory field while maintaining the resolution of the feature map, which effectively solves the problem of small target feature loss caused by the traditional downsampling operation. At the same time, the added P2 shallow feature extraction path complements the deep semantic features, enhancing the network's feature expression capability for multi-scale targets.

In terms of algorithm optimisation, this study adopts the WIoU v3 (Weighted IoU) loss function as an alternative to the traditional IoU metric, which enables the model to adaptively adjust the loss weights of samples with different difficulty levels through its dynamic focusing mechanism and outlier-based gradient gain assignment strategy, effectively enhancing the localisation accuracy of the detection frame. Compared with the loss function based on geometric feature constraints, WIoU v3 significantly enhances the detection robustness of tiny and densely distributed targets while maintaining computational efficiency. Experimental validation on the VisDrone2019 dataset shows that the improved algorithm strikes a good balance between detection accuracy and computational efficiency. In particular, the detection of small and densely distributed targets is significantly improved, while the computational complexity of the model is kept at a low level, demonstrating its practical application on UAV embedded platforms. These improvements provide new technical ideas for aerial target detection in complex scenes.

Acknowledgements

This paper is supported by Projects of major scientific and technological research of Ningbo City (2022Z090(2022Z050), 2023Z050(the second batch)), Projects of major scientific and technological research of Beilun District, Ningbo City(2021BLG002, 2022G009), Projects of scientific and technological research of colleges student's of China(202313022036, 202413001008).

Reference

[1] Uchinoura S ,Miyao J ,Kurita T .An Object Detection Method Using Probability Maps for Instance Segmentation to Mask Background:Regular Papers[J].Journal of Advanced Computational Intelligence and Intelligent Informatics,2023,27(5):886-895.

[2] Tang X ,Jia C ,He Z .UAV Path Planning: A Dual-Population Cooperative Honey Badger Algorithm for Staged Fusion of Multiple Differential Evolutionary Strategies[J].Biomimetics,2025,10(3):168-168.

[3] Leilei F ,Jun Y ,Zhiyi H .SSD Object Detection Algorithm Based on Feature Fusion and Channel Attention[J].International Journal of Advanced Network, Monitoring and Controls,2022,7(3):80-89.

[4] Su J ,Song Z ,Wang X , et al.Maize Seedling and Weed Detection Using BFSL-YOLOv8[J].World Scientific Research Journal,2025,11(3):20-28.

[5] Han J ,Chen H ,Ding Y , et al.You Only Look Once–Aluminum: A Detection Model for Complex Aluminum Surface Defects Based on Improved YOLOv8[J].Symmetry,2025,17(5):724-724.

[6] Wang X ,Xiang X .DBF-YOLO: a fall detection algorithm for complex scenes[J].Signal, Image and Video Processing,2025,19(7):532-532.

[7] Firdiantika M I ,Kim S .IS-YOLO: A YOLOv7-based Detection Method for Small Ship Detection in Infrared Images With Heterogeneous Backgrounds[J].International Journal of Control, Automation and Systems,2024,22(11):3295-3302.

[8] Zhao X ,Zhang H ,Zhang W , et al.MSUD-YOLO: A Novel Multiscale Small Object Detection Model for UAV Aerial Images[J].Drones,2025,9(6):429-429