IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 November 2023, accepted 10 January 2024, date of publication 12 January 2024, date of current version 22 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3353308

RESEARCH ARTICLE

Improved YOLOv5s Algorithm for Small Target Detection in UAV Aerial Photography

SHIXIN LI[1], CHEN LIU[1], KAIWEN TANG[2], FANRUN MENG[1], ZHIREN ZHU[1], LIMING ZHOU[1], AND FANKAI CHEN[1]

[1]School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China
[2]Fu Foundation School of Engineering and Applied Science, Columbia University, New York, NY 10027, USA

Corresponding author: Chen Liu (18722547456@163.com)

ABSTRACT UAV aerial photos tend to have complicated backgrounds and dense targets that vary in size. Applying existing object detection algorithms to such images is often inaccurate and prone to misdetection and omission. To better improve the detection performance of UAV aerial photography, we proposed an improved small-target detection algorithm based on YOLOv5s: 1) We reconstructed the feature fusion network by introducing an upsampling layer, increasing the model's focus on features from small targets and improving related detection accuracy. 2) We introduced the SPD convolutional building block to downsample the feature map without losing learning information, improving the model's feature extraction ability. 3) We replaced the CIoU Loss function of the original model with EIoU to reduce the location loss during training and improve the regression accuracy. We experimented with the improved algorithm on the VisDrone2019 dataset and achieved mAP@0.5 of 44%, demonstrating a 10.7% improvement from the original model. The detection speed also increases to 99 FPS, indicating that the improved algorithm can maintain its real-time performance while improving its accuracy.

INDEX TERMS YOLOv5s, UAV, SPD, small target detection, EIoU.

I. INTRODUCTION

With the development of UAV technology, target detection technology based on deep learning has become an important research topic in the field of UAV applications, realizing target detection and recognition of ground scenes under the aerial photography viewpoint. However, the background of UAV aerial images is complex, the detection object is primarily tiny and easily obscured, and the target scale varies significantly due to the influence of the aerial photography viewpoint, which brings many challenges to the target detection of UAVs. Conventional target detection algorithms, when applied to UAVs, tend to lack detection accuracy. Hence, optimizing target detection algorithms for UAVs is particularly important.Traditional target detection algorithms have low accuracy, poor detection efficiency, and insufficient generalization and robustness. Along with the rise of Convolutional Neural Networks (CNN), target detection algorithms based on deep learning have gradually replaced the traditional ones. Contemporarily, two main categories for deep-learning-based target detection algorithms exist: two-stage models represented by Faster-RCNN [1] and single-stage models depicted by the YOLO series and SSD [2]. Since the proposal of YOLO [3], it has been widely used for target detection in various scenarios due to its fast detection speed. However, its detection accuracy on the small targets remains to be improved.To improve the problem of poor detection performance of small targets in UAV aerial photography, the authors made a new attempt at the YOLOv5s algorithm. We proposed an improved algorithm based on YOLOv5s, which can effectively improve the detection accuracy of small targets under the premise of keeping the size of the model, the number of parameters, and the detection speed similar to that of the original model.

II. RELATED WORK

Based on the characteristics of UAV aerial images and the research difficulties at this stage, Liu et al. [4]. proposed a

The associate editor coordinating the review of this manuscript and approving it for publication was Amin Zehtabian.

feature enhancement block (FEBlock) and integrated it into spatial pyramid pooling (SPP) to generate enhanced spatial pyramid pooling to improve the feature extraction ability of YOLOv5s network.However, this method was unable to demonstrate significant improvement in detection accuracy. Small targets still suffer from misdetection and omission. Li et al. [5]. added a small target detection layer on the basis of YOLOv5, and introduced a bidirectional feature pyramid (BiFPN) to fuse feature information from different scales to enhance the feature extraction ability of small targets in the image. Zhu et al. [6] proposed TPH-YOLO: an improved UAV target detection algorithm with new detection heads to detect objects with smaller scales, replacing the original prediction heads with Transformer Prediction Heads (TPH). This new algorithm improved the model's detection accuracy for small targets. Gao et al. [7].introduced the convolutional block attention module (CBAM) in YOLOv5 and added a small target detection layer to solve the problem of semantic loss when detecting small targets. Liu et al. [8] added the efficient channel attention (ECA) module to the backbone network of YOLOv5l and replaced the sampling method with transposed convolution to retain more feature information of small targets. However, the improved model based on YOLOv5l has a large number of parameters, which is not favourable for deployment on edge devices such as UAVs.

The above UAV target detection algorithms improved the detection ability of small targets mainly by adding a small-target detection layer in the feature fusion network and incorporating an attention mechanism. However, increasing the number of detection layers will lead to an increase in model size, number of parameters, and computational complexity. The detection speed of the model will also be affected as a result. At the same time, due to the multiple downsampling operations in the CNN used, it is difficult to effectively retain features of small targets, leaving the problem of missing feature information of small targets in the detection process unattended in current algorithms.

To address the pitfalls of the existing algorithms, we designed an improved small-target detection algorithm for UAV aerial photography based on YOLOv5s. This algorithm can effectively improve the detection accuracy on small targets while keeping the model size, number of parameters, and detection speed like those of the original model. To solve the problem of poor detection accuracy of small targets, we reconstructed the feature fusion network of YOLOv5s by adding an upsampling layer to adjust the output scale of the feature maps, resulting in feature images with low receptive fields and high resolution. Adding this upsampling layer retains more features of small targets, and effectively improves their detection accuracy. For the problem of the loss of features about small targets, we introduced the SPD convolutional building block to replace the stride convolution in the original model. This new convolutional block downsamples the feature maps without losing information, which improves the feature extraction capability of the model. Finally, we replaced the CIoU Loss [9] of the original model with the EIoU Loss [10], reducing the location loss in the training time and improving the regression accuracy of the model.

III. THE YOLOV5S ALGORITHM

The YOLOv5s network consists of four parts: input, backbone, neck, and detection head.

(1) **Input:** The input images are preprocessed using the Mosaic augmentation to further improve the training speed and detection accuracy of the model. It also has adaptive anchor frame calculation and adaptive image scaling methods.

(2) **Backbone:** YOLOv5s uses CSPDarknet53 as the backbone network, fusing the CBS convolutional layers and C3 module for feature extraction from the input image. At the end of the backbone network is the Spatial Pyramid Pooling Fast (SPPF) module, which converts the feature maps into feature vectors of fixed size.

(3) **Neck:** The neck part consists of Feature Pyramid Networks (FPN) [11] and Path Aggregation Network (PAN) [12]. The combination of the two enhances the feature fusion capability of YOLOv5s.

(4) **Head:** The detection head performs convolution on three different sizes of feature maps outputted from the Neck for target category and location regression detection.

IV. THE IMPROVED YOLOV5S ALGORITHM

To better improve the detection accuracy of small targets in UAV aerial images, we improved YOLOv5s. The network structure of the improved model is shown in Fig. 1:

A. RECONFIGURING THE FEATURE FUSION NETWORK

The feature fusion network of YOLOv5s is composed of a Feature Pyramid Network (FPN) and a Path Aggregation Network (PANet). Figure 2 shows the schematic diagram of the FPN and PAN structures. In the process of shallow to deep feature extraction, the shallow features have higher resolution and richer geometric information, while the deep features have strong receptive field and rich semantic information. Most of the UAV aerial images contain small targets, which are heavily represented by shallow features. After multiple convolution and pooling operations, the original network outputs a feature map with low resolution and lacks the expression of shallow information, which makes it difficult for the original model to learn the features of the small targets, thus affecting the detection accuracy of small targets.

To improve the utilization of shallow features, we reconfigured the feature fusion network of the original model by adding an upsampling layer after the second up-sampling operation. This change allows the feature map to undergo the convolution and up-sampling operation again to adjust the output scale of the feature map, and ultimately obtain the feature maps with the sizes of 160 * 160, 80 * 80, and 40 * 40. Among them, the feature map of 160 *160 size is specialized

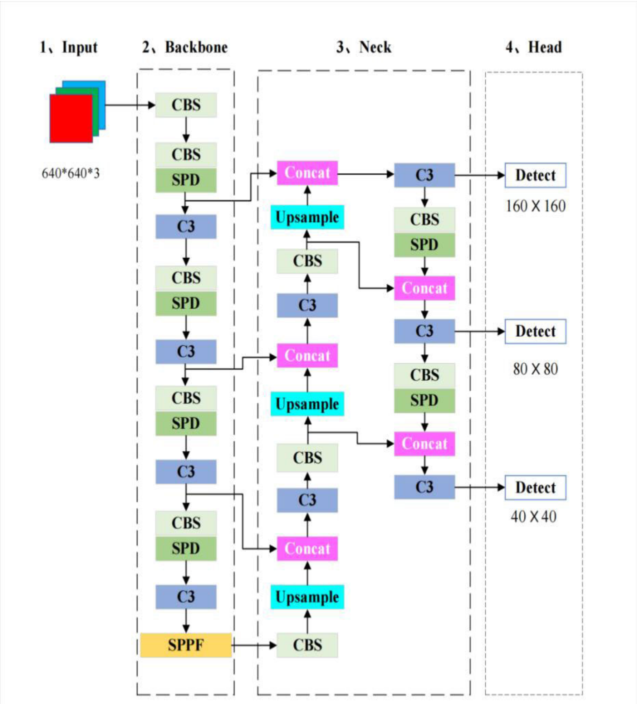

FIGURE 1. Improved YOLOv5s network structure.

in dealing with the detection of smaller objects in the image. It has higher resolution and better preservation of shallow feature information, making it easier for the network to learn the features of small targets, thus improving the detection accuracy of small targets.

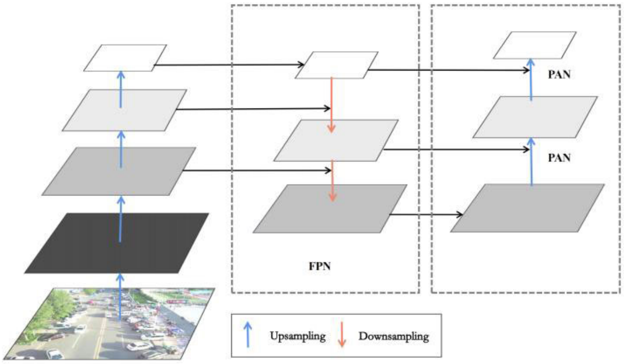

FIGURE 2. FPN+PAN schematic diagram.

B. THE INTRODUCTION OF THE SPD CONVOLUTION BUILDING BLOCK

The background of UAV aerial images is complex. The feature information of small targets is also vague. The original model utilizes a stride convolutional layer with a step size of 2 to downsample the feature map, which is likely to result in the lack of discriminative information in the process of feature extraction, thus affecting the accuracy of the detection. To better extract the feature information of the small target, we replaced the original convolutional structure with the SPD convolutional building block, so as to solve the problem of missing feature information due to the downsampling. This innovation can effectively improve the model's performance in detecting small targets.

The SPD Convolution building block consists of the Space-to-Depth layer followed by a non-strided convolution layer. Its working mechanism is shown in Figure 3:

(1) **Slicing:** Slices the original feature map into several sub-feature maps according to the scale factor. For example, when **scale** $= \mathbf{2}$, 4 sub-feature maps can be obtained. The size of each sub-feature map is $(\frac{S}{2}, \frac{S}{2}, C_1)$. This is equivalent to a 2-fold downsampling operation of the original feature map.

(2) **Concatenation:** Perform stitching operations on sub-feature maps by channel dimension to obtain intermediate feature maps X′. The spatial dimensions of the intermediate feature maps are downscaled by a factor of scale, while the channel dimensions are upcaled by a factor of scale. This will result in a final dimension of $(\frac{\mathrm{S}}{\mathrm{scale}}, \frac{\mathrm{S}}{\mathrm{scale}}, \mathrm{scale}^2 * \mathrm{C}_1)$

(3) **Non-strided Convolution:** The intermediate feature maps X' are further converted to the final feature maps X'' by using a non-strided (Stride =1) convolutional layer with a C_2 filter. The purpose of using a non-strided convolutional layer is to retain all the feature discriminative information as much as possible.

The original image is first sliced into sub-feature maps after the SPD convolutional building block. Then, the sub-feature maps are concatenated, and the features are extracted. Finally, the extracted feature information is filtered so that the feature information in the image is preserved to the maximum extent while the feature map is downsampled. The feature extraction capability of the network is improved after this process.

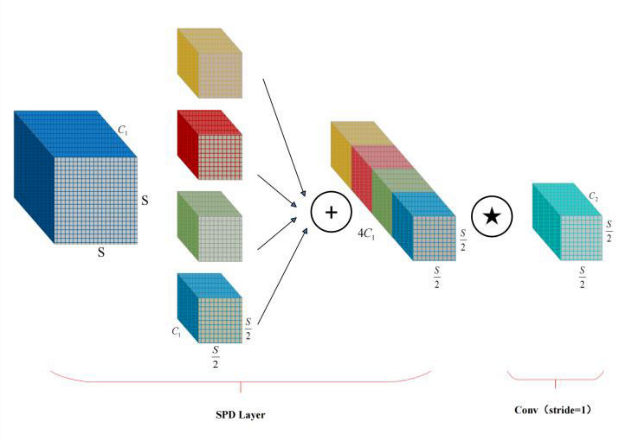

FIGURE 3. Schematic diagram of SPD principle.

C. IMPROVED LOSS FUNCTION

The original YOLOv5s model uses CIoU as the location loss function. CIoU considers three important geometric factors: overlap area, distance between centers, and aspect ratio. Given the prediction bounding boxes B and the ground truth bounding boxes B^{gt}, the CIoU loss function is defined as follows:

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

where b and b^{gt} represent the center of B and B^{gt} ρ^2 represents the Euclidian distance. c is the diagonal length of the smallest outer bounding box of the predicted and real bounding boxes. α is the positive equilibrium parameter. v is the aspect ratio of the predicted and real bounding boxes.

Since the last parameter, v, in L_{CIoU} only reflects the difference in aspect ratio but not the actual difference between width and height respectively with their confidence levels, it will hinder the optimization of the model similarity. Based on the above analysis, this paper adopts the EIoU loss function to replace the CIoU loss function in the original network model, with the aim of obtaining better localization results while speeding up the convergence of the model. The EIoU loss function is defined as follows:

$$\begin{aligned} L_{EIoU} &\\ &= L_{IoU} + L_{dis} + L_{asp} \\ &= 1 - IoU + \frac{p^2(b, b^{gt})}{(w^c)^2 + (h^c)^2} + \frac{p^2(w, w^{gt})}{(w^c)^2} + \frac{p^2(h, h^{gt})}{(h^c)^2}, \end{aligned} \tag{2}$$

where $\boldsymbol{h^w}$ and $\boldsymbol{h^c}$ is the width and height of the smallest outer bounding box of the prediction and target bounding boxes.

EIoU divides the loss function into three parts: the IoU loss $\mathrm{L}_{\mathrm{IoU}}$, the distance loss $\mathrm{L}_{\mathrm{dis}}$, and the aspect ratio loss $\mathrm{L}_{\mathrm{asp}}$, which takes into account the overlapping area, the distance between the center points, and the real difference of the length and width of the edge. It also solves the fuzzy definition of the aspect ratio based on CIoU. The loss term of the aspect ratio is split into the difference between the width and height of the prediction bounding box and the width and height of the minimum outer bounding box, which accelerates the convergence speed of the model and improves the regression accuracy.

V. EXPERIMENTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTINGS

We conducted the experiment on a machine with an i7-11700 CPU, an NVIDIA RTX A4000 GPU, 16GB of video memory, and operates on Windows 10. Pytorch1.13.1 was used to run the code, and the training was accelerated using CUDA11.7.1. The training parameters are set as: batch size = 16, number of epochs = 300, and other settings are default.

B. EXPERIMENTAL DATASET

We selected VisDrone2019 as the dataset for experiments in this paper. This dataset is collected by the AISKYEYE team at the Machine Learning and Data Mining Laboratory at Tianjin University. Images in this dataset are captured by a variety of UAV cameras. They cover a wide range of locations (from 14 different cities thousands of kilometers apart in China), environments (urban and rural), objects (pedestrians, vehicles, bicycles, etc.), and densities (sparse and congested scenes). The VisDrone2019 dataset contains 10 categories of detected targets, and as can be seen in Fig. 4, most of the objects in the dataset have a size less than 0.1 times the size of the original image, which is in line with the definition of the relative scale of small targets and can meet the needs of experimental validation. The VisDrone2019 dataset contains 6,471 images in the training dataset, 548 images in the validation dataset, and 1,580 images in the test dataset.

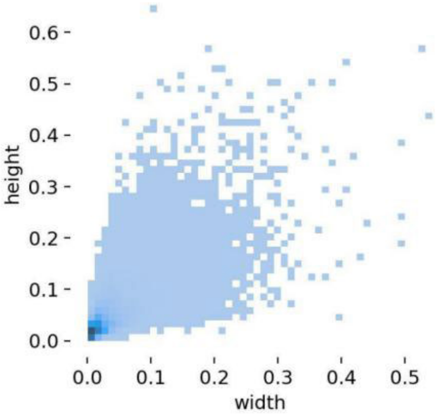

FIGURE 4. Target size distribution of the VisDrone dataset.

C. EVALUATION CRITERIA

The experiments use mean Average Precision (mAP), Precision (P), Recall (R), Frames Per Second (FPS), and the number of parameters (Params) to evaluate the detection performance of the model.

(1) **mAP** denotes the mean average precision of all detection categories and is calculated as:

$$mAP = \frac{1}{c}\sum_{i=1}^{c} AP_i \tag{3}$$

where c represents the number of total detection classes. i represents the number of detections. AP is the average precision of a single class. mAP is obtained by averaging the APs of all categories. mAP@0.5 represents the mean average precision when the threshold IoU is set as 0.5. mAP@0.5:0.95 represents the mean average precisions when the threshold IoU is set as each value between 0.5 and 0.95 with a stride of 0.05.

(2) **Precision (P)** indicates how accurately the model detects the target. Its formula is:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{4}$$

where TP represents the number of positive samples that have been predicted as positive, *FP* represents the number of positive samples that have been predicted wrong as negative.

(3) **Recall (R)** measures of the model correctly identifying True Positives.

$$R = \frac{TP}{TP + FN} \times 100\% \tag{5}$$

(4) **Frames Per Second (FPS)** denotes the refresh rate of the model inference speed and is calculated as follows: Use either SI (MKS) or CGS as primary units. (SI units are strongly encouraged.) English units may be used as

$$FPS = \frac{Framenum}{ElapsedTime} \tag{6}$$

where Framenum is the total processed frames, ElapsedTime is the total time taken to process the frames.

(5) **Params** represent the number of parameters occupied by the model memory (unit: M)

D. THE ABLATION EXPERIMENT WITH THE IMPROVED MODEL

We created seven sets of ablation experiments to verify the enhancement of model detection performance by each improved module. The validation was carried out by sequentially adding each module to the baseline model YOLOv5s, and the results of the ablation experiments are shown in Table 1:

TABLE 1. Ablation experiment.

	EIoU	*SPD*	*Upsampling*	*mAP@0.5*	*mAP@0.5:0.95*
ex1	-	-	-	*0.333*	*0.179*
ex2	√	-	-	*0.341*	*0.182*
ex3	-	√	-	*0.355*	*0.193*
ex4	-	-	√	*0.41*	*0.232*
ex5	√	√	-	*0.367*	*0.199*
ex6	-	√	√	*0.434*	*0.249*
ex7	√	-	√	*0.42*	*0.237*
ex8	√	√	√	*0.44*	*0.254*

Experiment one is conducted using the original model YOLOv5s. Experiment two replaced the loss function with EIoU, resulting in a 0.8% increase in mAP@0.5. Experiment three introduced the SPD convolutional building block, increasing mAP@0.5 by 2.2%. Experiment four added one upsampling layer for the feature fusion network and increased mAP@0.5 by 7.7%. This is because after adding one layer of upsampling, the output feature map has a smaller receptive field and is rich in shallow information, which is conducive to feature extraction and fusion of small targets.

Experiments 5, 6, and 7 are a combination of the three improvement points, which all improved the model's performance compared to the original version and showed a superimposed effect on the performance enhancement. Experiment 8 is the final improved model that combines the three improvement points. It demonstrates the most significant improvement compared to the original model, increasing mAP@0.5 by 10.7% and mAP@0.5:0.95 by 7.5%. Recall (R) is improved by 9.3 percent.

E. EXPERIMENTAL ANALYSIS OF THE UPSAMPLING MODULE

To verify the effect of adding an upsampling layer on the detection performance of the model, we conducted comparison experiments with the YOLOv5s model with an added layer of upsampling and the original YOLOv5s. The experimental data are shown in Table 2. The detection accuracy of each class is significantly improved after adding one layer of upsampling. Among them, the accuracy of detecting pedestrians rises by 9.6%. The accuracy of detecting cars rises by 8.5%, and the accuracy of detecting buses rises by 11.2%. These are the three classes with the biggest rise in detection accuracy. mAP@0.5 increased by 7.7 percent compared to the original model. It is clear that adding a layer of up-sampling to the feature fusion network can effectively improve the detection accuracy of the model for small targets.

TABLE 2. Comparison of detection accuracy after increasing upsampling.

Detection category	*YOLOV5s*	*Add a layer of upsampling*
pedestrian	*0.405*	*0.501*
people	*0.319*	*0.38*
bicycle	*0.103*	*0.173*
car	*0.738*	*0.823*
van	*0.355*	*0.445*
truck	*0.283*	*0.354*
tricycle	*0.199*	*0.275*
awning-tricycle	*0.107*	*0.138*
bus	*0.43*	*0.542*
motor	*0.391*	*0.47*
all	*0.333*	*0.41*

F. COMPARISON OF DIFFERENT LOSS FUNCTIONS

In order to verify the effect of different loss functions on the performance of the model, we selected GIoU Loss [13], EIoU Loss, and SIoU Loss [14] for comparison experiments with CIoU Loss in the original model. The experimental results are shown in Table 3. After replacing the loss function with EIoU, mAP@0.5 increased by 0.8%. This loss function has the best effect on the model detection performance compared to the other three. The EIoU loss function can reduce the location loss during model training and can effectively improve the regression accuracy of the model. As can be seen from Fig. 5, replacing the CIoU with EIoU accelerates the convergence speed of the model while improving the detection accuracy.

G. COMPARISONS WITH DIFFERENT DOWNSAMPLED CONVOLUTIONS

In order to verify the effectiveness of SPD convolutional building blocks for model performance improvement, we respectively replace all strided convolutions with strid = 2 in the original model with Omni-Dimensional Dynamic Convolution(OD-Conv) [15], Depthwise

TABLE 3. Comparison of loss function.

Loss function	*mAP@0.5*	*mAP@0.5:0.95*
CIoU	*0.333*	*0.179*
GIoU	*0.321*	*0.173*
SIoU	*0.334*	*0.178*
EIoU	*0.341*	*0.182*

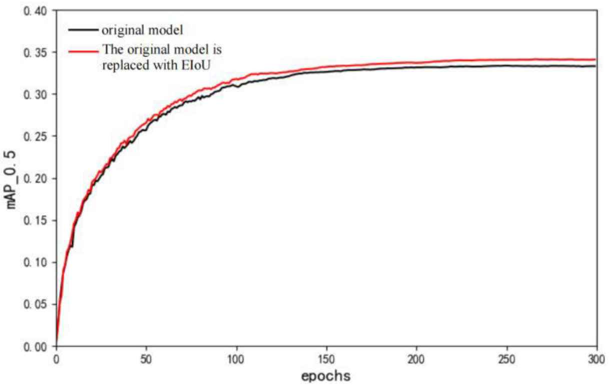

FIGURE 5. Performance Comparison of Replacing EIoU.

Overparameterized Convolution (DO-Conv) [16], and SPD convolutional building blocks to conduct comparative experiments. The results as shown in Table 4. As can be seen from the table, after using OD-Conv to replace the strided convolution from the original model, the average precision, accuracy, and recall have decreased compared with the original model. DO-Conv is the one with the least number of parameters among the four convolutions and the one with the poorest performance index. After using the SPD convolution building block to replace the original model's stride convolution, the number of params rises only slightly compared to the original model. However, mAP@0.5 increased by 2.2%. It can be seen that the introduction of SPD convolution has a substantial effect on the improvement of model detection performance.

TABLE 4. Comparison of different downsampling convolutions.

Different convolutional layers	*mAP@0.5*	*P*	*R*	*Params(M)*
Conv	*0.333*	*0.445*	*0.338*	*14.4*
ODConv	*0.332*	*0.468*	*0.332*	*14.5*
DO-Conv	*0.283*	*0.41*	*0.292*	*10.1*
SPD	*0.355*	*0.504*	*0.343*	*17.5*

H. COMPREHENSIVE COMPARISONS

Finally, to verify the detection performance of the improved model in this paper more comprehensively, we compared the performance of our model with other state-of-the-art models. We used the following criterias: mAP@0.5, precision, recall, model size, and parameter size. The results are shown in Table 5. From these data, we can conclude that the improved model, compared with other mainstream models, has the highest average precision, accuracy, and recall. Compared with the baseline model, our model increased mAP@0.5 by 10.7% only at the cost of increasing the number of fewer parameters. Our model has an FPS of 99, which may satisfy the requirements of real-time detections.

TABLE 5. Performance comparison of different algorithms.

algorithm	*mAP @0.5*	*P*	*R*	*FPS*	*Params (M)*	*Model size(MB)*
YOLOv7-tiny[17]	*0.358*	*0.488*	*0.37*	*125*	*6.0*	*12.3*
YOLOv8s	*0.398*	*0.508*	*0.388*	*222.2*	*11.1*	*22.5*
TPH-YOLOv5	*0.394*	*0.496*	*0.4*	*60.2*	*41.6*	*83.6*
YOLOv5m	*0.37*	*0.467*	*0.382*	*86.2*	*20.9*	*42.2*
YOLOv5l	*0.395*	*0.52*	*0.387*	*80.6*	*46.2*	*92.9*
YOLOv5s	*0.333*	*0.445*	*0.338*	*108.7*	*7.0*	*14.4*
ours	***0.44***	***0.592***	***0.431***	***99***	***8.6***	***17.9***

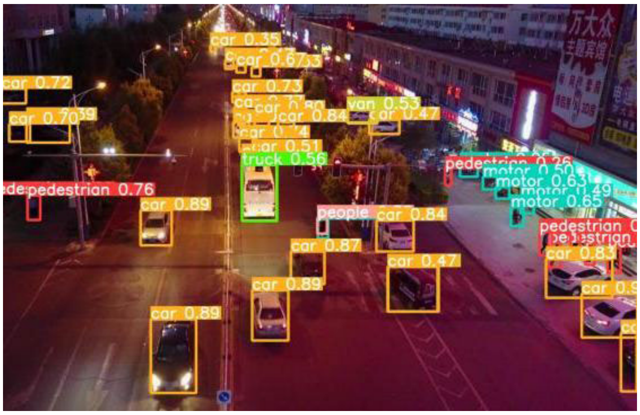

FIGURE 6. Night City Road(Original YOLOv5s).

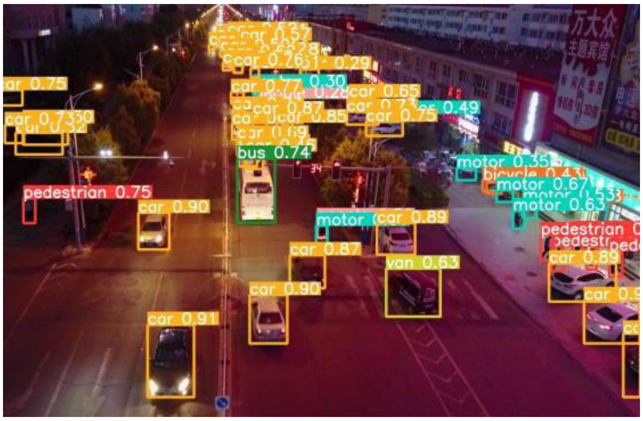

FIGURE 7. Night City Road(Improved Algorithm).

I. COMPREHENSIVE VISUALIZATION

In order to more intuitively compare the detection performance of the improved model with the original model, we selected UAV aerial images from three scenes, namely, city roads at night, complex traffic intersections, and town

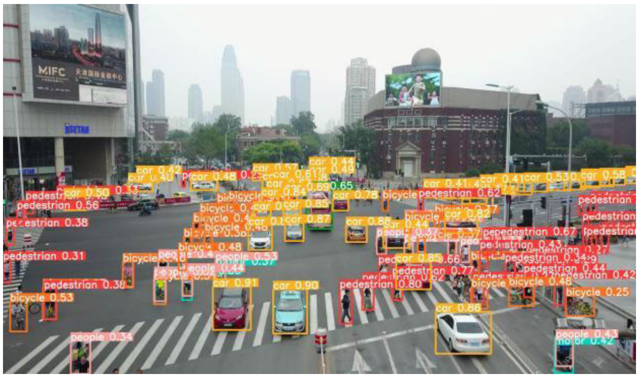

FIGURE 8. Complex Traffic Intersection(Original YOLOv5s).

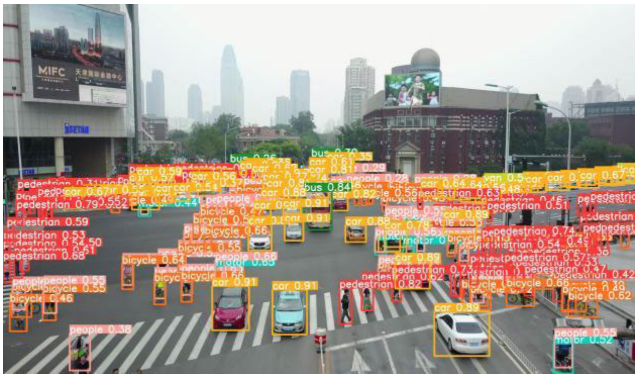

FIGURE 9. Complex Traffic Intersection(Improved Algorithm).

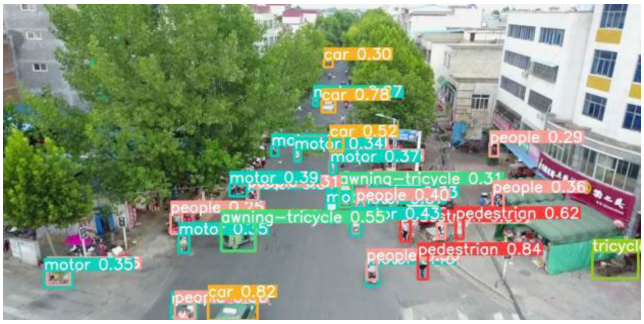

FIGURE 10. City Road with Shades(Original YOLOv5s).

boulevards, for experiments. These images are characterized by insufficient light, dense small targets, complex background, and severe occlusion, which can verify the detection effect of the model under extreme conditions. In Fig. 6 and Fig 7, it can be seen that the original algorithm falsely detects buses as trucks and vans as cars. However, the improved algorithm can detect them correctly. It also detected targets farther away. In Fig. 8 and Fig 9, the improved algorithm detects more densely crowded targets than the original algorithm. Small targets farther away can be recognized effectively. In Fig. 10 and Fig 11, the improved algorithm is able to detect the occluded targets, which the original algorithm fails to detect.

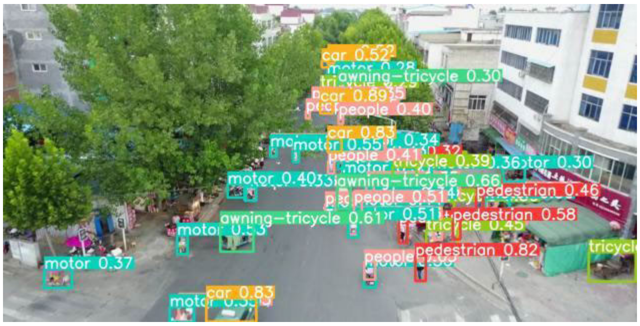

FIGURE 11. City Road with Shades(Improved Algorithm).

VI. CONCLUSION

To address the poor performance of current algorithms for detecting small targets in UAV aerial photography, we designed an improved small target detection algorithm based on YOLOv5s. By adding an upsampling layer to the feature fusion network of YOLOv5s, we managed to narrow the receptive field and improve the resolution of feature images. The network now pays attention to more feature information of small targets, improving its detection accuracy for small targets. The SPD convolutional building block is used to replace the stride convolution in the original model to perform the downsampling operation, retaining the feature discriminative information to the greatest extent. This improves the missing feature information caused by multiple downsampling of ordinary stride convolution and improves the feature extraction ability of the model for small targets. The original loss function is replaced with EIoU to reduce the location loss during training and improve the accuracy of the model. Results from the experiments show that the improved model has a 10.7% increase in its mAP@0.5, reaching 44%. While improving its accuracy, the improved model can still meet the needs of real-time detections. Compared with other mainstream algorithms, the improved model has better detection performance for small targets in UAV aerial photography. Targeted optimization of the network structure will be continued in subsequent research to reduce the number of parameters to achieve a lightweight model for better application and deployment on edge devices with limited computational power.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.

[4] Z. Liu, X. Gao, Y. Wan, J. Wang, and H. Lyu, "An improved YOLOv5 method for small object detection in UAV capture scenes," *IEEE Access*, vol. 11, pp. 14365–14374, 2023.

[5] S. Li, X. Yang, X. Lin, Y. Zhang, and J. Wu, "Real-time vehicle detection from UAV aerial images based on improved YOLOv5," *Sensors*, vol. 23, no. 12, p. 5634, Jun. 2023.

[6] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.

[7] T. Gao, M. Wushouer, and G. Tuerhong, "Small object detection method based on improved YOLOv5," in *Proc. Int. Conf. Virtual Reality, Human-Computer Interact. Artif. Intell. (VRHCIAI)*, Oct. 2022, pp. 144–149.

[8] S. Liu, P. Liang, Y. Duan, Y. Zhang, and J. Feng, "Small target detection for unmanned aerial vehicle images based on YOLOv5l," in *Proc. 10th Int. Conf. Inf. Syst. Comput. Technol. (ISCTech)*, Dec. 2022, pp. 210–214.

[9] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," 2020, *arXiv:2005.03572*.

[10] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," 2021, *arXiv:2101.08158*.

[11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Washington, DC, USA: IEEE Computer Society, Jul. 2017, pp. 936–944.

[12] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8439–8448.

[13] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[14] Z. Gevorgyan, "SIoU Loss: More powerful learning for bounding box regression," 2022, *arXiv:2205.12740*.

[15] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," in *Proc. Int. Conf. Learn. Represent.*, vol. 1181, 2022, pp. 1–20.

[16] J. Cao, Y. Li, M. Sun, Y. Chen, D. Lischinski, D. Cohen-Or, B. Chen, and C. Tu, "DO-Conv: Depthwise over-parameterized convolutional layer," 2020, *arXiv:2006.12030*.

[17] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

SHIXIN LI received the master's degree and the Ph.D. degree in photonics from Tianjin University, China, in 2000 and 2003, respectively. From 2003 to 2005, he was a Postdoctoral Student with the Institute of Electronics, Chinese Academy of Sciences. From 2005 to 2012, he was a Senior Engineer with the Tianjin Institute of Navigational Instruments. He is currently a Professor with the Tianjin University of Technology and Education, Tianjin, China. His research interests include the artificial intelligence and intelligent information processing. He has published approximately 60 research articles in these areas.

CHEN LIU was born in Hulunbuir, Inner Mongolia, China, in 2000. He received the bachelor's degree from the College of Electronic Engineering, Tianjin University of Technology and Education, in 2021, where he is currently pursuing the master's degree. From 2017 to 2021, he mainly researched the development and application of embedded circuits. After 2021, his research direction is computer vision, mainly studying object detection algorithms based on deep learning. He has published several articles and participated in many research projects in this field.

KAIWEN TANG received the Bachelor of Science degree (Hons.) in computer software engineering from the University of Alberta, in 2023. He is currently pursuing the master's degree in computer science, specializing in natural language processing with Columbia University. He was a Student Software Developer with Calian, AT. He was mainly responsible for the development of the satellite uplink system for a major satellite radio company in North America. He also has one-year of working experience with the Open AI Laboratory, Nanjing, specializing in software development and computer vision.

FANRUN MENG was born in Linfen, Shanxi, China, in 1999. He received the B.S. degree from the School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin, China, in 2021, where he is currently pursuing the M.S. degree. From 2018 to 2022, he mainly researches embedded development and electro magnetic compatibility (EMC). After 2022, he began to study artificial intelligence algorithms, mainly on path planning and deep learning-based target detection algorithms. He has also publishing several articles in this field.

ZHIREN ZHU was born in Qingyang, Gansu, China, in 1999. She received the bachelor's degree from the College of Electronic Engineering, Tianjin University of Technology and Education, in 2021, where she is currently pursuing the master's degree. From 2018 to 2021, she mainly studied digital circuit technology, during which multiple circuit design was completed. In 2022, she has studied artificial intelligence algorithms, mainly studying target detection algorithms based on deep learning. She has published many articles in this field.

LIMING ZHOU was born in Suihua, Heilongjiang, China, in 2000. She received the bachelor's degree from the College of Electronic Engineering, Tianjin University of Technology and Education, in 2022, where she is currently pursuing the master's degree. From 2019 to 2022, she mainly researches on embedded development and participates in the development of many embedded application devices. After 2022, she began to research on artificial intelligence algorithms, mainly on target detection algorithms based on deep learning. She has also published several research articles in this field.

FANKAI CHEN was born in Chizhou, Anhui, China, in 1998. He received the bachelor's degree from the College of Electronic Engineering, Tianjin University of Technology and Education, in 2021, where he is currently pursuing the master's degree. From 2018 to 2021, he mainly researches the direction of the industrial IoT and participates in the development of a number of the industrial IoT intelligent devices. After 2021, he began to research on artificial intelligence algorithms, mainly on path planning algorithms based on bionic algorithms and target detection algorithms based on deep learning. He has also published several research articles in this field.