SPD-CBAM: Small Target Detection in UAV Images

Haorui Huang

{HRHUANG2026@outlook.com}

College of Information Engineering, Shanghai Maritime University, Shanghai, China

Abstract. To address the issues of low pixel ratio of extremely small targets and the susceptibility to missed detections and false detections in traditional detection algorithms under complex environmental conditions during drone aerial surveillance, this paper proposes a new detection scheme based on YOLOv8s.This study first addresses the issue of small target feature loss caused by stride convolution during downsampling by introducing a spatial-to-depth (SPD) transformation mechanism to reconstruct the backbone network. By changing the feature mapping method, key detail information is preserved. In the feature fusion stage, a convolutional block attention module (CBAM) is specifically embedded to enhance the target response weights from both channel and spatial dimensions, thereby reducing the interference of complex background noise. Experimental results show that the improved model achieves an mAP@50 of 46.4%, a significant improvement of 15.9% compared to the original algorithm, and a significant increase in recall of 29.9%.

Keywords: UAV vision; Small target detection; Spatial-to-depth convolution (SPD-Conv); CBAM.

1 Introduction

As unmanned aerial vehicle (UAV) technology develops quickly, these platforms equipped with high-performance visual sensors have been widely used in key areas such as traffic flow monitoring, power line inspection, forest fire prevention and emergency search and rescue. In these applications, automatic target detection technology is the core of realizing intelligent tasks. However, due to the high flight altitude and varied perspective of UAVs, the targets captured (such as pedestrians and vehicles) occupy very few pixels in the image. According to statistics from authoritative datasets such as VisDrone[1], a large number of targets occupy only 8×8 pixels or even smaller. Such targets are defined as "micro-targets" in academia. Achieving accurate capture of micro-targets can not only improve the perception accuracy of UAV systems but also has profound practical significance for ensuring public safety and improving search and rescue efficiency.

At present, single-stage target detectors represented by YOLOv8 have become the preferred choice for UAV platforms due to their excellent inference speed. However, since they generally use strided convolution for feature map downsampling, they will cause information

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365346

loss. Although strided convolution could expand the receptive field and reduce computational cost, for small targets with a very low pixel ratio, this "jump" sampling will ignore a lot of detailed information, which is an irreversible information loss process, resulting in the discriminative features of weak targets being severely attenuated before entering the deep network. In addition, there is another major problem: messy textures, shadows and occlusions in aerial images are often confused with target features, and a single feature extraction mode is difficult to accurately lock effective targets in complex semantic backgrounds [2].

To address the "feature evaporation" problem in the downsampling process, this paper constructs a detection framework that can balance feature preservation and background suppression. The core starting point of the research is to use YOLOv8s[3] developed by Ultralytics as the benchmark model, reshape the information flow logic in the downsampling process, and use the space-to-depth (SPD) conversion mechanism to replace the original lossy sampling, thus establishing a lossless feature transfer path. Meanwhile, a multivariate attention management mechanism is used to enable dynamic focusing of features in complex backgrounds, thereby radically resolving the issue of substantial missed detection of extremely tiny targets.

The innovative research work in this paper is mainly reflected in three aspects:

To tackle the challenging problem of feature loss when considering drones, a backbone network founded on SPD-Conv was devised. This network achieves pixel - level mapping from the spatial domain to the channel domain and retains the fundamental texture details of small targets to the greatest possible degree.

In the feature fusion stage, a convolutional block attention module (CBAM) is introduced. Through the combined effect of channel and spatial attention, the model's background recognition ability in cluttered environments is significantly improved.

The algorithm was validated on the VisDrone-2019 dataset. Experimental results indicate that the algorithm achieves significant improvements over the benchmark model in terms of mAP@50 and recall, providing a feasible technical path for deploying high-precision perception systems on the UAV edge.

2 Improved detection algorithm design

2.1 Overall network architecture

The SPD-CBAM network architecture proposed in this paper is based on YOLOv8s and is deeply customized for the specific characteristics of the UAV perspective. The overall structure is built based on the design principles of "lossless feature transfer" and "selective feature enhancement".

The improved model consists of three key parts. The first part is the backbone network, which replaces all the original strided convolutional layers with SPD-Conv modules. This is done by changing the physical arrangement of the data, transforming the spatial details that would otherwise be lost during downsampling into channel features, thereby providing richer underlying semantics for subsequent layers.

The second part is that a CBAM attention module is embedded after the C2f module in the feature path of the PAN-FPN structure [4][5]. The function of the CBAM attention module is to perform fine filtering on the redundant features transmitted from the backbone network. These redundant features are generated due to the sharp increase in the number of channels.

The third part is the detection head retaining the old decoupled head design which ensures that classification and localization tasks can be optimized independently.

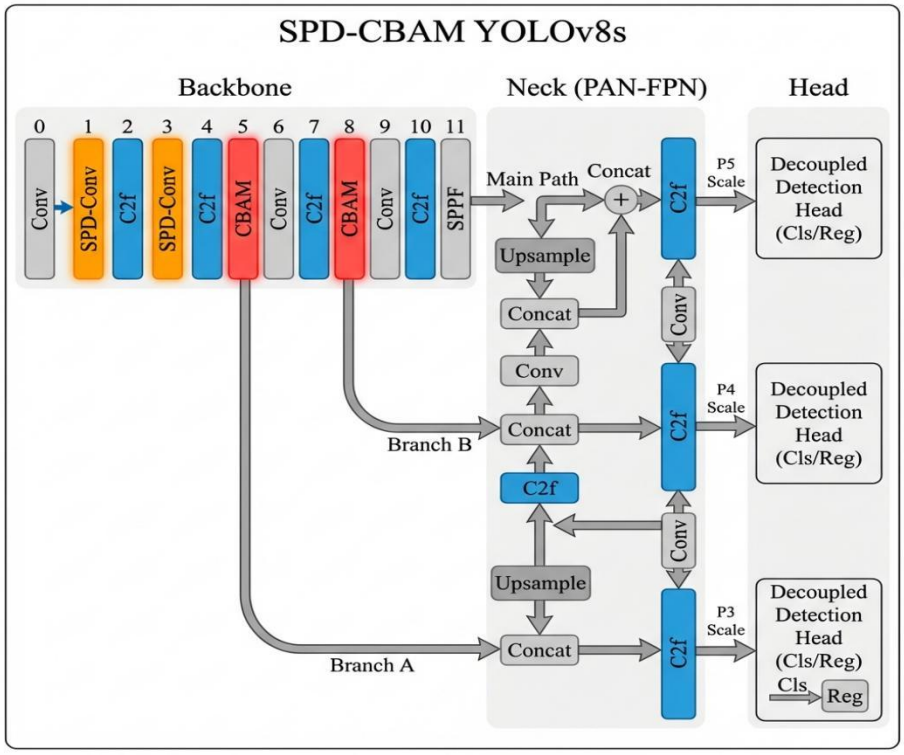

Fig. 1. The Overall Architecture of YOLOv8s with SPD-Conv and CBAM Enhancements (Picture credit: Original)

Standard strided convolutions discard spatial information during downsampling, making it difficult to detect tiny objects in UAV imagery. To preserve these details, we integrated Space-to-Depth (SPD-Conv) modules into the shallow layers of the backbone (Blocks 1 and 3, Figure 1). Rather than dropping pixels, SPD-Conv reshapes spatial blocks into the channel dimension, converting an $S \times S \times C$ feature map into $S/2 \times S/2 \times 4C$. This approach reduces spatial resolution while retaining the original data. Keeping this raw pixel information intact early in the network ensures that deeper layers have the low-level features necessary to locate small targets.

Beyond scale variation, UAV images typically contain cluttered backgrounds. We placed Convolutional Block Attention Modules (CBAM) at Blocks 5 and 8 to filter this noise before the fusion stage. CBAM sequentially calculates attention across two dimensions: channel attention amplifies target-relevant semantics, while spatial attention isolates target locations.

As Figure 1 illustrates, the network routes these refined features directly into the PAN-FPN neck via Branch A and Branch B. Injecting attention-filtered data directly into the multi-scale fusion process prevents background noise from dominating the feature maps, which improves overall detection accuracy.

2.2 Lossless downsampling based on spatial-to-depth transformation

Traditional downsampling is essentially a form of spatial information loss. For low-pixel targets in the VisDrone dataset, after two traditional downsampling operations, the target will shrink or even disappear on the feature map [6].

The SPD module maps spatial information to the channel dimension through pixel-level recombination. This operation can theoretically be traced back to the sub-pixel convolution proposed by Shi al. [7]. Although the application scenarios are different, both utilize the interchangeability between spatial resolution and channel depth to achieve lossless feature transfer. This paper draws on this idea to preserve the original distribution features of extremely low pixel targets during the downsampling stage.

Traditional downsampling can cause the target to disappear, while SPD can avoid this situation. This article explains the principle of SPD [8]. As shown in Figure 2, the SPD layer is implemented by "pixel cutting and stacking" the input tensor. Given an input feature map $X\,(S, S,\ C_1)$, the SPD module divides it into a series of sub-tensors and concatenates them along the channel dimension.

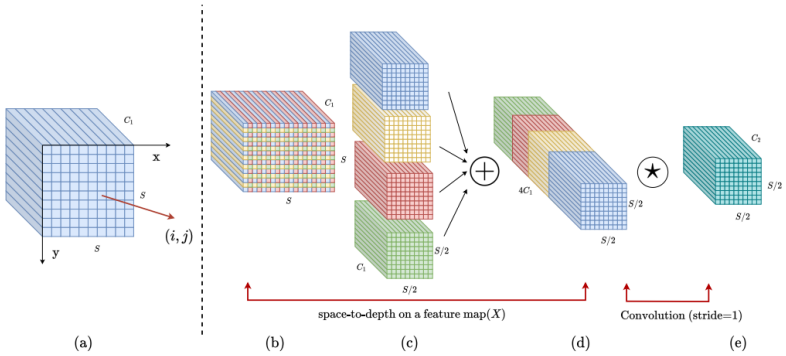

Fig. 2. The schematic diagram of the pixel rearrangement and downsampling principle of the SPD-Conv module [8]

Assuming a downsampling ratio of $s = 2$, for input coordinates(i, j), the transformed channel index mapping relationship can be expressed as:

$$f_{x,y}(i, j) = X(2i + x, 2j + y, c), \quad x, y \in 0,1 \tag{1}$$

At this point, the feature map size changes from (H, W, C) to$(\frac{H}{2}, \frac{W}{2}, 4C)$.

This transformation does not involve any pixel sampling; instead, it converts the spatial resolution loss into an increase in channel-dimensional information. This allows the model to still "retrieve" the original pixel distribution features even in deeper network layers.

At the current stage, the dimension of the feature map changes its shape from (H, W, C) into $(\frac{H}{2}, \frac{W}{2}, 4C)$. This converting process does not have any pixel sampling step; On the contrary, it results in a loss of spatial resolution into the increment of information on the channel dimension. This permits the model to still "obtain back" the original pixel distribution characteristics even in deeper network layers.

2.3 Convolutional Block Attention Enhancement Module

For the reason that unmanned aerial vehicle aerial pictures are complex and include a great deal of redundant information, Lai Qinbo and other scholars put forward to combine attention mechanism and dilated convolution to strengthen the model's capability of perceiving targets[9]. Under the inspiration of this thought, this paper brings in CBAM at the feature fusion phase, and further restrains the disturbance of complicated background noise to the identification of small targets via feature reconstruction in channel and spatial dimensions.

After bringing in SPD-Conv, the channel dimension of the feature map has an obvious enlargement. Although this method can retain the information, it also brings very much background noise and calculation expenditure. This current article introduces the CBAM module [10] to rebuild the feature flow in two processing stages.

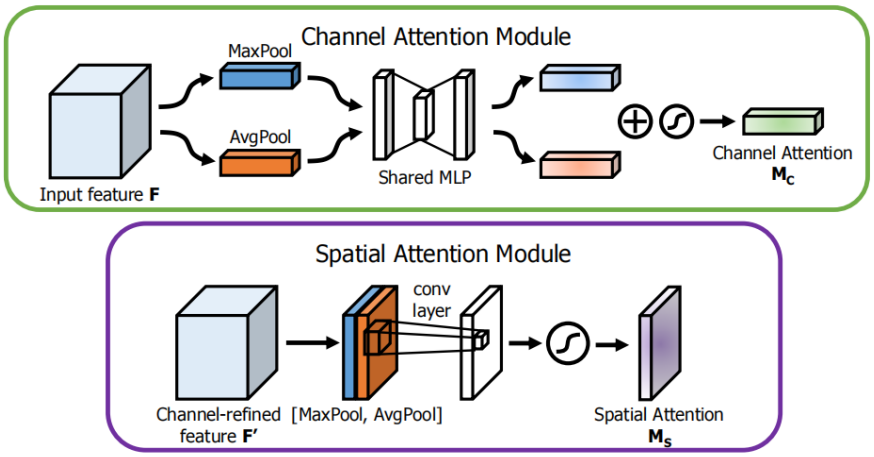

Fig. 3. Diagram of each attention sub-module [10]

Just like what Figure 3 displays, the internal structure of CBAM contains channel attention and spatial attention: Channel attention uses global average pooling (GAP) and global max pooling (GMP) in the parallel way to get channel features, and its mathematical expression is written below:

$$M_c(F) = \sigma\left(MLP\left(AvgPool(F)\right) + MLP\left(MaxPool(F)\right)\right) \tag{2}$$

These steps build the connection among channels and give bigger weights to channels that hold "human" traits.

The spatial attention mechanism carries out compression on the channel axis after channel refinement, hence it generates a spatial mask by utilizing 7×7 convolution.

$$M_s(F) = \sigma\left(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])\right) \quad (3)$$

In Formulas (2) and (3),$F \in \mathbb{R}^{C\times H\times W}$ represents the input feature map. $M_c \in \mathbb{R}^{\times 1\times 1}$ and

$M_s \in \mathbb{R}^{1\times H\times W}$ denote the channel and spatial attention maps, respectively. σ refers to the sigmoid activation function. $MLP$is a shared multi-layer perceptron with one hidden layer. AvgPool and MaxPool indicate global average pooling and global max pooling operations. In the spatial attention formula, $f^{7\times7}$ signifies a convolution operation with a 7×7 kernel size, and $[\cdot;\cdot]$ represents the concatenation operation along the channel axis.

This step addresses the question of "where is the target?" by suppressing the response of non-target areas (such as road edges and trees), thereby enhancing the model's resilience against cluttered backgrounds.

3 Experimental Content

3.1 Experimental Dataset

The experimental data come from the authoritative evaluation benchmark VisDrone-DET2019 from the perspective of UAVs [11]. In the data preparation stage, considering the complexity of the original VisDrone label system, this paper selects the cleaned and format-aligned visdrone-xionq-g0wdk version from the Roboflow platform [12] for eliminating the interference of label noise on detector training.

This dataset version, while preserving the original image semantics, underwent standardized label format conversion for typical targets in drone aerial photography. The experiment involved 8626 images, covering key categories such as pedestrians and vehicles. Through the platform's integrated management, the consistency of data partitioning (70% of samples for training, 20% for validation, and 10% for testing) was ensured, providing a standardized benchmark environment for subsequent experiments.

3.2 Experimental environment and parameter configuration

The experiment was conducted using the PyTorch framework in the Colab environment with an NVIDIA Tesla T4/A100 processor. Specific parameters are shown in Table 1.

Table 1. Experimental Environment and Hyperparameter Settings

Classification	Parameters	Values
Dataset related	Dataset	VisDrone-DET2019
	Target Classes	Pedestrian & People
	Split	Train (6037) / Val (1726) / Test (863)
	Ratio	7:2:1
Hardware environment	Hardware	NVIDIA Tesla T4 / A100 (Google Colab)
	Acceleration	CUDA 11.x / 12.x
Software Architecture	Framework	PyTorch 2.x
	Baseline	Ultralytics YOLOv8s

Hyperparameter configuration	Total Epochs	150 (30 Warm-up + 120 Fine-tuning)
	Image Size	640×640
	Batch Size	16
	Optimizer	SGD
	lr0	0.01
	Weight Decay	0.0005
Data Augmentation	Augmentation	Mosaic (0.8), Flip (0.5), Mixup (0.15)

3.3 Experimental Results and Analysis

Table 2. Performance Comparison Analysis of the Improved Model and the Baseline Model

Index	Baseline	Improved model	Absolute improvement (percentage points)	Relative improvement rate (%)
mAP@50	30.5%	46.4%	+15.9%	+52.1%
Precision	54.2%	63.4%	+9.2%	+16.9%
Recall	32.8%	42.6%	+9.8%	+29.9%

According to what Table 2 displays, the ameliorated model that this paper puts forward obtains prominent enhancement on all the core evaluation measurement indices.

At the beginning stage, the average accuracy (mAP@50) has exhibited obvious promotion. The model's mAP@50 accuracy has increased from 30. Five percent to forty-six. 4 percent, an absolute growth of 15. 9 percentage points, and there is a relative increase of 52. 1%. This promotion of performance carries out the quantification of the efficiency of the added module when it extracts key features from the complex scenes.

The analysis that was carried out afterwards on detection accuracy showed that the accuracy had an increase from 54. Two percent to sixty-three. Four percent, a relative growth of 16. 9%. This outcome tells us that the promoted model is able to hold back background noise more accurately and reduce the possibility of wrong alarms.

From the angle of unmanned aerial vehicles, the work of recalling has all along been a work with many difficulties in the detection of small objects. Experiment outcomes prove that the recall ratio increased from 32. 8 percent to 42.6 percent, a 29.9 percent growth. This obvious promotion shows that through promoting the feature extraction network (e. g., which brings in SPD-Conv for decreasing information loss) and making the attention mechanism become stronger (e. g., which uses CBAM to rebuild the prominent features), the model's capability of capturing blocked and tiny objects has been promoted, thus greatly relieving the problem of detection missing.

The improved model presented in this paper achieves increased recall while maintaining high accuracy. In challenging tasks like VisDrone, the mAP improvement reaches 52.1%, demonstrating the effectiveness and practical value of the proposed algorithm optimization scheme.

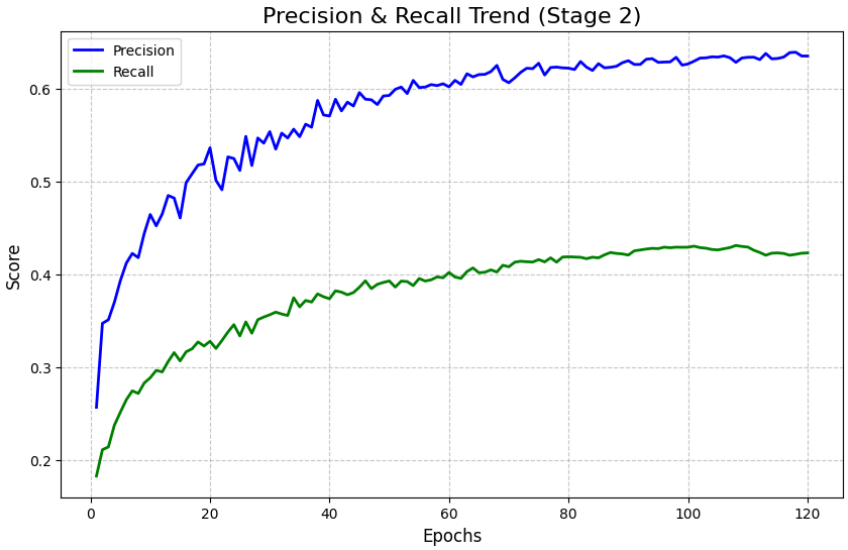

Fig. 4. PR curve of the improved model on a single pedestrian category (Picture credit: Original)

Figure 4 shows the precision-recall curve (P - R) which is got by this experiment. The relation between the area that is under the curve and the coordinate axis expresses the mAP@50 indicator for the pedestrian category. According to the data which is in Table 2, the model has obtained an mAP of 46. 4 percentage points for one individual category, that is, pedestrian. When the curve extends to the right along the recall axis, the precision does not have a very sharp drop. This curve form, which keeps high recall and at the same time shows a steady downward trend, is hard for the traditional YOLOv8s to reach when they process VisDrone data. This phenomenon gives indication that although the bringing in of SPD-Conv enlarges the channel features (it is helpful for finding more pedestrians that were missed before, thus recall gets improvement by 29. 9%), the following CBAM attention mechanism has the function of "cleaning", it effectively suppresses the complex background noise which exists in the image, hence therefore it ensures the accuracy of this experiment conduct.

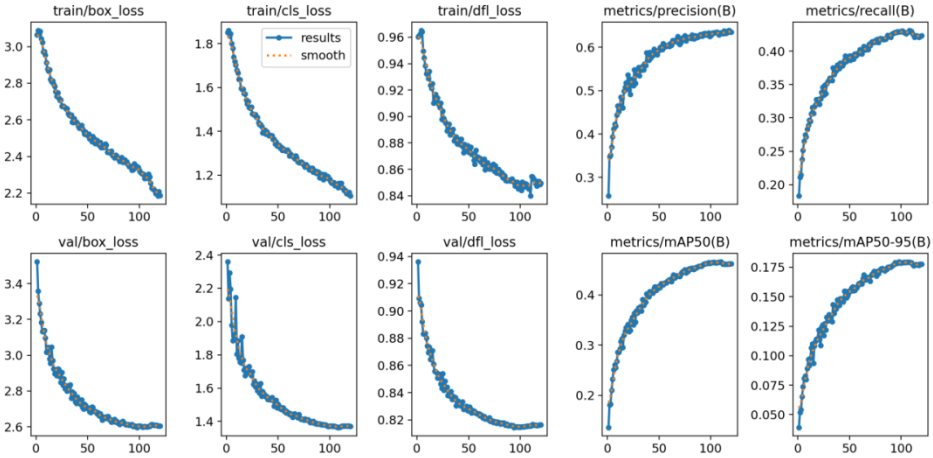

Fig. 5. Loss and performance evaluation curves during model training (Picture credit: Original)

Figure 5 has shown the target detection model's performance indexes after 120 training rounds of promotion. The loss curve is frequently utilized by people to measure the discrepancy that lies between the model's predicted values and the real values; In general, the smaller this numerical value is, the better the outcome hence is. All three measurement indexes in the training collection present a stable and continuously decreasing trend, and the three measurement indexes in the checking collection also decrease and gradually get closer to a stable condition. The loss curve of the verification set very closely coincides with the curve of the training set, having no fluctuations or rebounds. This indicates that the model does not have obvious overfitting.

The precision of this model in the end became stable at 0. 641. This score is what can reflect the model's accuracy in the identification of positive samples. When put beside the baseline model that was said before, the false positive ratio has a great reduction; therefore, this shows feature extraction has made progress. The recall rate has obtained stabilization at about the value 0.435. In view of the complexity that the detection task possesses, this numerical value indicates that the model is able to capture the majority of target features, hence it demonstrates that the model has strong capability for object detection.

4 Conclusions

This article puts emphasis on the difficulties of finding small objects in complicated backdrops and closely packed things, in particular those which are connected with unmanned aerial vehicles (UAVs). One algorithm of YOLOv8 that has been improved has been put forward and carried out. This ameliorated method decreases feature information loss in the procedure of downsampling through the integration of the SPD-Conv module, and therefore enhances the model's ability for reconstructing key features by the incorporation of the CBAM attention mechanism. This paper has been carried out many times of deep training and verification on the VisDrone dataset, and therefore, has obtained satisfying results.

The monitoring data which have collected indicates that the model obtains strong convergence on both the training set and the validation set, with decreases being present in the numerical values of multiple kinds of loss functions. In the end, this model obtains an accuracy value which is 0. six four one and an mAP of 0. 5 of 0. 468, which represents a quite obvious promotion compared with the baseline model. These outcomes prove that the ameliorated model efficaciously strengthens the capacity for recognizing tiny targets, hence keeping high accuracy all the while. This current work gives effective technical support for promoting the accuracy of UAV image detection, hence enabling more accurate distinguishing in actual scenes such as traffic monitoring and disaster relief, and thus providing feasible optimization schemes for visual tasks under complex visual angles.

Although the positioning exactness has been promoted to a certain degree, the recall ratio is zero. 435, which shows that improvement space still exists, it means that challenges still remain in recalling targets that are extremely small. The future research work will place emphasis on further developing multi-scale feature fusion mechanisms, and therefore explore the lightweight deployment of this model on edge devices, hence to realize more efficient and all-round real-time detection in complex and dynamic environments.

References

[1] Du, D., Zhu, P., Wen, L., et al.: VisDrone-DET2019: The vision meets drone object detection in image challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0-0 (2019)

[2] Tong, K., Wu, Y., Zhou, F.: Recent advances in small object detection based on deep learning: A review. Image and Vision Computing. 97, pp. 103910 (2020)

[3] Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8. (2023). https://github.com/ultralytics/ultralytics

[4] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, pp. 936-944 (2017)

[5] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path Aggregation Network for Instance Segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, pp. 8759-8768 (2018)

[6] Sunkara, R., Luo, T.: No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), pp. 1-16 (2022)

[7] Shi, W., Caballero, J., Huszár, F., et al.: Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In: CVPR (2016)

[8] Sunkara, R., Luo, T.: No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In: ECML PKDD, pp. 1-16 (2022)

[9] Lai, Q., Ma, Z., Zhu, R.: UAV IMAGE OBJECT DETECTION BASED ON ATTENTION MECHANISM AND DILATED CONVOLUTION. Computer Applications and Software. 42(2), pp. 227-235 (2025)

[10] Woo, S., Park, J., Lee, J.Y., et al.: CBAM: Convolutional Block Attention Module. In: European Conference on Computer Vision (ECCV). Springer, Cham, pp. 3-19 (2018)

[11] Zhu, P., Wen, L., Bian, X., Ling, H., Hu, Q.: Vision Meets Drones: A Challenge. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2018)
[12] hwx: visdrone-xionq-g0wdk Dataset. Roboflow (2024). https://app.roboflow.com/hwx/visdrone-xionq-g0wdk/1