IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 January 2023, accepted 19 January 2023, date of publication 31 January 2023, date of current version 15 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241005

RESEARCH ARTICLE

An Improved YOLOv5 Method for Small Object Detection in UAV Capture Scenes

ZHEN LIU, XUEHUI GAO, (Member, IEEE), YU WAN, JIANHAO WANG, AND HAO LYU
College of Intelligent Equipment, Shandong University of Science and Technology, Tai'an, Shandong 271002, China

Corresponding author: Xuehui Gao (xhgao@163.com)

This work was supported in part by the Tai'an Science and Technology Development Program under Grant 2021GX037, and in part by the National Natural Science Foundation of China under Grant 62073198.

ABSTRACT Aiming at the problem of a large number of small dense objects in high-altitude shooting and complex background noise interference in the captured scenes, an improved object detection algorithm for YOLOv5 UAV capture scenes is proposed. A Feature Enhancement Block (FEBlock) is first proposed to generate adaptive weights for different receptive field features by convolution, assigning major weights to shallow feature maps to improve small object feature extraction ability. The FEBlock is then integrated into Spatial Pyramid Pooling (SPP) to generate Enhanced Spatial Pyramid Pooling (ESPP), which performs feature enhancement for the result of each maximum pooling; and creates new features containing multi-scale contextual information with better feature characterization capability by weighting fused contextual features. Secondly, the Self-Characteristic Expansion Plate (SCEP) is proposed, which achieves the fusion and expansion of feature information through compression, non-linear mapping, and expansion with its own module, further improving the network's capacity for feature extraction and generating a new spatial pyramid pooling (ESPP-S) by splicing with ESPP. Finally, a shallower feature map is added as a detection layer to the YOLOv5 network model's large, medium, and small detection layers to improve the network's detection performance for medium and long-range objects. Experiments were conducted on the VisDrone2021 dataset, and the results showed that the improved YOLOv5 model improved mAP0.5 by 4.6%, mAP0.5:0.95 by 2.9%, and precision by 2.7%. The mAP0.5 of the model trained at the input resolution of 1024 × 1024 reached 56.8%. The experiments show that the improved YOLOv5 model can improve object detection accuracy for UAV capture scenes.

INDEX TERMS Feature enhancement, small object detection, UAV, YOLOv5.

I. INTRODUCTION

As UAV technology continues to evolve, camera-equipped UAVs or general-purpose drones have been rapidly deployed for various applications, including agriculture, aerial photography, public safety, ecological protection, and more. Therefore, the requirements for an intuitive understanding of visual data collected from these platforms are getting higher and higher. Object detection technology based on deep learning is more and more closely applied to UAVs. However, the high altitude at which UAVs fly, the large number of small-sized objects in the captured images, and the complex background noise interference between small dense objects lead to a significant decrease in detection accuracy [1]. This makes it difficult to detect objects in UAV capture scenes, so it is important to design a method to improve the detection accuracy of small objects in images.

Deep learning techniques have advanced quickly in recent years, and numerous Convolutional Neural Network (CNN)-based object detection algorithms have been proposed and used to detect objects in UAV images. Two main types of object detection algorithms exist two-stage-based and single-stage-based methods. Object detection based on the two-stage method is also known as the candidate region-based method. Firstly, the candidate box is extracted according to the image, and then the detection point result is obtained by secondary correction based on the candidate region. The detection accuracy is high, but the detection speed is slow. The first work

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

of this kind of algorithm is RCNN (Region CNN) [2], then Fast-RCNN (Fast Region-Based CNN) [3] and Faster-RCNN (Faster Region-Based CNN) [4] improved it in turn. Due to its excellent performance, Faster-RCNN is still a very competitive algorithm in the field of object detection. Subsequently, algorithms such as FPN (Feature Pyramid Network) [5] and Mask RCNN [6] have proposed improvements to address the shortcomings of Faster RCNN, which further enriches the components of Faster-RCNN and enhances its performance. Compared to two-stage object detection algorithms, single-stage object a priori algorithms generate detection results by computing directly on the image, with fast detection low speed but lower detection accuracy. The pioneer of this type of algorithm is YOLO (You Only Look Once) [7]. Subsequently, SSD (Single Shot MultiBox Detector) [8] and Retinanet [9] improved it in turn, and the subsequent improved versions YOLOv2 [10], YOLOv3 [11], YOLOv4 [12] and YOLOv5 based on YOLO. Although the prediction accuracy is less than the two-stage object detection algorithm, YOLO can detect UAV images due to its all-around performance.

More specifically, scholars have extensively researched object detection for UAV capture scenes. The literature [13] combines the Spatial Attention Module (SAM) with Channel Attention Module (CAM), improves the fully connected layer after feature compression in CAM, and changes the connection structure of SAM and CAM, thus proposing a spatial-channel attention module (SCAM) and using it on YOLOv5 to improve spatial dimensional feature capture, which not only reduces the computational effort but also improves to some extent the accuracy. The literature [14] proposed TPH-YOLOv5, which added a prediction head to YOLOv5 and applied a Transformer Encoder Block to the head part to form Transformer Prediction Heads (TPH), which improved the detection of high-density occluded objects in UAV images. In literature [15], a Scale Selection Pyramid Network (SSPNet) for minutiae detection was proposed by using the Context Attention Module (CAM), Scale Enhancement Module (SEM), and Scale Selection Module (SSM) to suppress the gradient computation inconsistency problem in FPN by controlling the data flow of adjacent layers. To solve the problem of false detection and missed detection caused by occlusion conditions, literature [16] improved the generalization ability of the detection network through data cleaning and enhancement, and set a priori anchor frame, and adjusted the confidence loss function of the detection layer based on IoU (Intersection over Union) to reconstruct the network. The literature [17] uses a bidirectional feature pyramid network for necking and introduces a SimAM attention module to fuse features effectively. The literature [18] proposes a new detection network, DCLANet, to crop and locally attend to dense small people in UAV images to solve the problem that the network cannot focus on small objects. In summary, deep learning methods have high application value in UAV image object detection, and a lot of results have been achieved. However, further research is still needed to improve detection accuracy.

To further improve the object detection accuracy of UAV capture scenes and solve the problem of poor detection effect caused by too dense between small size object and object. In this paper, the Feature Enhancement Block (FEBlock) and the Self-Characteristic Expansion Plate (SCEP) are designed and introduced into the original Spatial Pyramid Pooling (SPP) [19] module of YOLOv5. The FEBlock is first embedded into the SPP, and then continues to fuse and expand the feature information through the SCEP module. A spatial pyramid pooling module ESPP-S with enhanced feature representation is proposed. In addition, a shallower feature map is added as a small object detection layer to improve the detection performance of the network for medium and long-range objects.

- The Feature Enhancement Block (FEBlock) is designed to enhance the receptive field and enable efficient fusion of different receptive field features. The FEBlock is also embedded in the SPP module to generate the Enhanced Spatial Pyramid Pooling (ESPP) module, which has stronger feature characterization capability than the original SPP module.
- The Self-Characteristic Expansion Plate (SCEP) is designed to realize the fusion and expansion of feature information. The ESPP module is spliced with the SCEP module, and the ESPP-S module is proposed, which can improve the small object detection ability.
- Based on the large, medium, and small detection layers of the YOLOv5 network model, a shallower feature map is added as the detection layer according to the dataset's characteristics to improve the detection performance of the network for medium and long-distance objects.

II. SMALL OBJECT DETECTION FOR UAV CAPTURE SCENE

This paper takes YOLOv5 version 6.1 as the benchmark network and makes subsequent improvements. YOLOv5 has five models: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Since the detection effect of YOLOv5x is better than the other four models, even if the calculation cost is higher than other models, we still choose YOLOv5x to pursue high detection performance.

A. YOLOv5 NETWORK STRUCTURE AND IMPROVEMENTS

The YOLOv5 algorithm is simple to deploy and train and has great dependability and stability. At the same time, it is also one of the most accurate single-stage object detection algorithms. Therefore, YOLOv5 is chosen in this study for further improvement and as the object detection algorithm for UAV capture scenes. YOLOv5 follows the detection idea of the YOLO series, that is, dividing the grid on the input image. When there is a center point of the detection object in a grid, the grid is used to detect the object. Four components make up the YOLOv5 model: Input, Backbone, Neck, and Head. Firstly, the image to be detected is processed by the input end and sent to the backbone network, and then the preliminary feature extraction is performed by CBS, C3, and SPPF (SPP-Fast). The backbone network generates feature maps of different sizes, and then enhances the ability to detect objects of different scales through PANet (Path Aggregation Network) [20]. Finally, three feature maps P3, P4, and P5, are generated to detect small, medium, and large objects in the picture. The Prediction Head uses a preset prior bounding box to perform confidence calculation and bounding box regression on each pixel in the three feature maps to obtain a multidimensional array including object class, class confidence, box coordinates, and width and height information. By setting the corresponding threshold to filter the useless information in the array and performing the non-maximum suppression (NMS) process, the final detection information can be output [21], [22].

For the problem of a large number of small dense objects in the UAV capture scenes and the presence of complex background noise interference, this paper proposes an improved small object detection algorithm based on YOLOv5. Figure 1 shows the structure of the improved YOLOv5 model. The overall network architecture optimizes the original network design from three aspects. The red dashed box shows the improved spatial pyramidal pooling. Firstly, the receptive field is increased by designing the feature enhancement block to improve the degree of attention to the small object area. The adaptive weights are formed for different receptive fields to improve the extraction ability of the model at different scales. The feature enhancement block is fused into SPP, and an Enhanced Spatial Pyramid Pooling (ESPP) module is proposed, which performs feature enhancement for the result of each maximum pooling, and generates new features containing multi-scale contextual information by weighting the fused contextual features. The feature enhancement block is introduced into SPP to improve the global feature extraction capability by weakening the background noise interference. Secondly, the feature information is further fused and expanded after stitching with a Self-Characteristic Expansion Plate. This gives the model better robustness and improves the detection capability of small dense objects. The blue dashed box represents a micro-scale detection layer that was created by collecting lower spatial features and combining them with high-level semantic features to improve the model's capacity to detect smaller objects.

B. IMPROVEMENTS IN SPATIAL PYRAMIDAL POOLING

The latest version of YOLOv5 uses SPPF, which replaces the three parallel max pooling in SPP with serial and modifies the pooling core size all to the same size. By streamlining the process of pooling, the duplication of SPP operations is avoided, and the speed of the network operation is improved. A comparison of the structure compared to the original SPP is shown in Figure 2.

Although SPPF speeds up the network's detection rate, the detection accuracy is not ideal when facing small dense objects. Therefore, this paper proposes a novel Spatial Pyramid Pooling (ESPP-S) based on SPP, which has stronger feature characterization capability than SPP and SPPF.

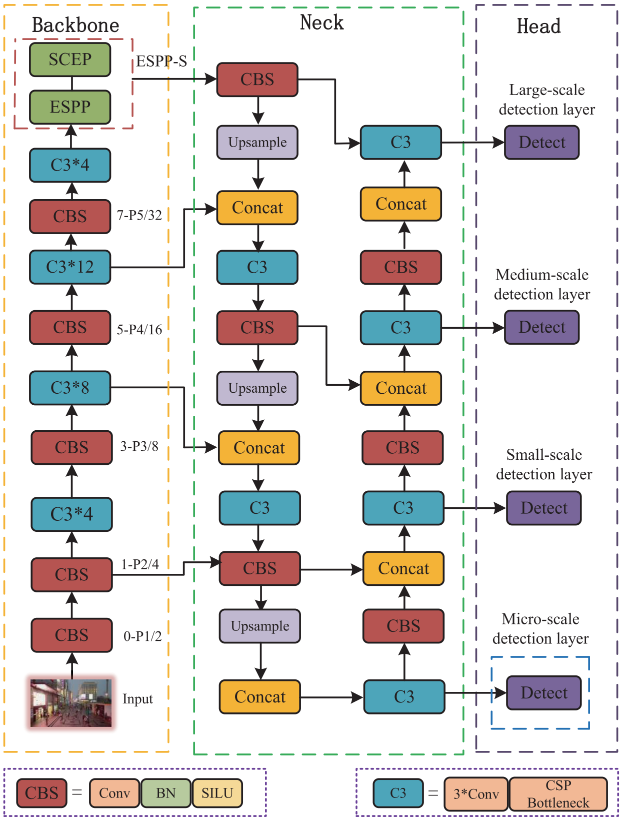
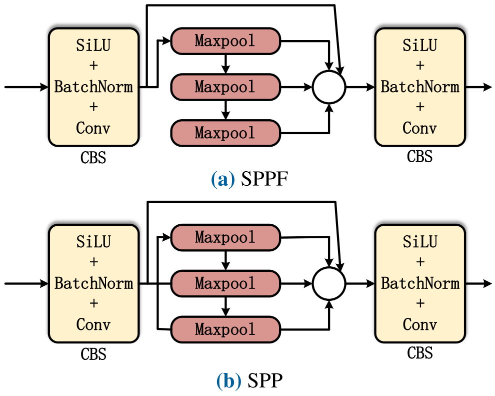


FIGURE 1. Structure of the improved YOLOv5 model.



(a) SPPF

(b) SPP

FIGURE 2. Comparison of SPPF and SPP structures.

Firstly, a Feature Enhancement Block (FEBlock) is designed to enhance the representation ability of features. In addition, the Self-Characteristic Expansion Plate (SCEP) is designed for feature information fusion and expansion. The feature enhancement block FEBlock is integrated into the spatial pyramid pooling module to generate enhanced spatial pyramid pooling (ESPP). Then the ESPP is spliced with the self-feature expansion block to generate ESPP-S, which improves the detection effect of small dense objects.

1) FEATURE ENHANCEMENT BLOCK (FEBlock)

FEBlock can be regarded as a feature enhancement block, which integrates information without deepening the network

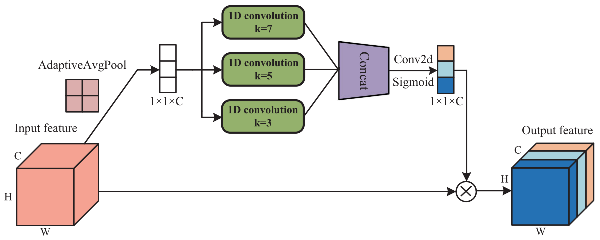


FIGURE 3. FEBlock structure diagram.

structure by combining different channel information. The structure uses an ECA-like attention mechanism to generate adaptive weights for different receptive field features through convolution, enabling the efficient fusion of different receptive field features and enhancing feature representation [23], [24]. The FEBlock is shown in Figure 3.

The features are first compressed into 1 × 1 scalars along the spatial dimension, and the output is shaped as a 1 × 1×C feature map by global averaging pooling, representing the global distribution of feature channel response values. A one-dimensional convolution follows this with three different convolution kernel sizes. The one-dimensional convolution acts as a non-fully connected layer, with each convolution acting on only some of the channels, allowing for full integration of some of the channel interactions through parallelism. This allows for proper cross-channel interaction and avoids the complexity of the model that a fully connected layer would otherwise create. Finally, the weights of each channel generated after the 2D convolutional feature transformation and Sigmoid feature mapping are multiplied by their respective weights [25].

Conventional convolution fuses all channels of the input feature map, and the network cannot focus on important feature channels. In contrast, FEBlock can adjust the distribution of weights, enhance useful features and suppress useless information [26]. When different scale feature maps are input, the model can adaptively adjust the size of the receiving domain of the small target in the UAV capture scene to improve the object detection performance of the model.

We introduce FEBlock into YOLOv5 as a feature enhancement module. Considering that the spatial pyramid module in YOLOv5 generates different scale contextual feature maps through pooling operations, the new ESPP module is obtained by introducing a feature enhancement module to the original spatial pyramid module to generate adaptive weights for different scale feature maps, and its structure is shown in Figure 4. The ESPP module first generates feature maps of different receptive fields through fixed-scale pooling branches, then compresses the channels through FEBlock, embeds spatial information into spatial attention maps, and generates new features containing multi-scale context information by weighted fusion context features. When feature maps of different scales are input, the model can adaptively adjust the size of the image's object acceptance region to highlight the feature map's object-related areas. Therefore, the ESPP module has a more robust feature representation ability than the original SPP module.

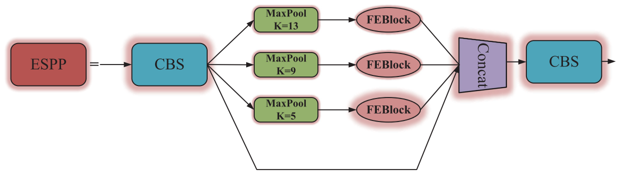


FIGURE 4. ESPP model structure.

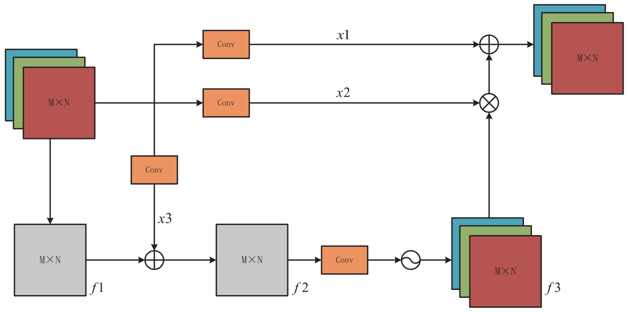


FIGURE 5. Structure of the SCEP model.

2) SELF-CHARACTERISTIC EXPANSION PLATE (SCEP)

The Self-Characteristic Expansion Plate extracts sub-block information on the feature map. The subgraph information is compressed, and the low-dimensional feature map (the high and low dimensions of this part refer to the number of channels of the feature map) is fused without scaling the number of channels of the original feature image. At the same time, the self-information is summarized with the original information after the sigmoid, which is conducive to the expansion of feature features. The fusion and expansion of feature information are realized through compression, nonlinear mapping, and expansion with its module. The structure of SCEP is shown in Figure 5.

SCEP first performs two-dimensional convolution on the input, extracts sub-block features on the feature map, and outputs $x1$, $x2$, and $x3$, respectively. The $x3$ is fused with the low-dimensional feature map for feature fusion, while feature extraction by convolution and Sigmoid feature mapping is aggregated with the original information to generate feature map $f3$, which facilitates feature expansion. The $x2$ is multiplied by $f2$ and then added to $x1$ to get the output. The process can be explained by equation (1).

$$\begin{aligned} f2 &= f1 + x3 \\ f3 &= f2 + Conv + Sigmoid \\ Output &= x1 + x2 \times f3 \end{aligned} \tag{1}$$

3) ESPP-S

The Feature Enhancement Block and the Self-Characteristic Expansion Plate are used in the ESPP-S network. Firstly, the Feature Enhancement Block is integrated into the spatial pyramid pooling to form the ESPP module. Then the ESPP module is connected with the Self-Characteristic Expansion

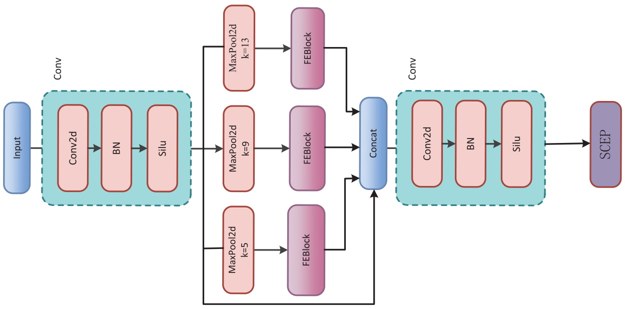


FIGURE 6. **Structure of the ESPP-S model.**

Plate to form the ESPP-S.The structure of ESPP-S is shown in Figure 6.

ESPP-S first performs convolution, BN and SiLU activation functions on the input image data. Convolution mainly performs deep feature extraction. BN allows the data to be processed before the activation function without causing unstable network performance due to oversized data. The activation function allows the nonlinearity of the network to be increased to fit various nonlinear functions. The next step is to go through a spatial pyramid pooling operation, consisting of three parallel kernels of 5, 9, and 13 for the maximum pooling, where the output results of the upper levels are pooled separately. The maximum pooling is done without changing the input data channels, and the input parameters change the width and height of the input data.

Feature enhancement is performed on the results of each maximum pooling. FEBlock first performs global average pooling on the input data, that is, the data with maximum pooling on the upper layer. The single-channel elements are compressed into 1 × 1 scalars, and the output 1 × 1×C feature map is used as the global spatial information on the feature channel. Three parallel one-dimensional convolutions then achieve the cross-channel interaction. The dimensionality of the data produced by these operations is transformed to match the dimensionality of the original data. Finally, the original input data is multiplied by the information flowing through each network layer. The maximum pooling data in each layer of the SPP network are respectively passed through the FEBlock module and spliced in the channel direction. The two dimensions of the final output length and width of the three shunts are consistent. Feature enhancement was performed on feature maps at different scales, allowing the network to focus more on small object regions, suppressing interference from background noise, and improving feature characterization. The feature maps are then streamed through the SCEP module. After compression, non-linear mapping, and expansion with its module, feature information fusion and expansion are achieved, further improving the detection of small objects.

C. ADDING A SMALL OBJECT DETECTION LAYER

Head outputs the prediction results, and the prediction includes the bounding box loss function and non-maximum suppression [27]. YOLOv5 uses the GIOU loss function as the bounding box loss function [21], and the GIOU is calculated as shown in equation (2)–(3). Assuming that A and B are any two properties, find a minimum closed shape C such that C can contain A, B. Then calculate the ratio of the area of C that does not cover A and B to the total area of C, subtracting this ratio from the IOU of A and B. It can better reflect the intersection of the predicted box and the real box, improving the speed and accuracy of detection. The three feature maps generated by the neck (dimensions 80 × 80, 40 × 40, and 20 × 20) are sent to the prediction head. Then a confidence calculation and bounding box regression are performed for each pixel in the feature map using a pre-defined prior anchor. A non-maximum suppression process is performed by setting the corresponding thresholds. However, these three layers of the feature map no longer meet the current detection needs, so improvements are made to the original ones.

$$IOU = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

$$GIOU = IOU - \frac{|C \backslash (A \cup B)|}{|C|} \tag{3}$$

This paper focuses on object detection in UAV capture scenarios, where the YOLOv5 algorithm is not ideal mainly due to the size of small objects. The feature maps are extracted by simple downsampling, which can lead to loss of information of interest when the down-sampling multiplier is too large, and forward propagation of the network when the downsampling multiplier is too small requires a large number of feature maps to be stored in memory, increasing the GPU resource usage and causing the training and inference to exceed the video memory. To avoid these problems and improve the accuracy and robustness of the network, according to the principle of the receptive field, based on the three detection layers of the original YOLOv5 head, this paper adds a shallower feature map as the detection layer according to the characteristics of the dataset. After the original detection layer, convolution and upsampling operations are added to expand the feature map further. Then, the obtained feature map is merged with the feature map extracted by the network backbone in order to get a bigger feature map for small object detection, which makes the network more sensitive to small objects under high-resolution images and enhances the network's ability to detect medium and long-range objects. In this paper, P2, P3, P4, and P5 four-layer feature maps are used to achieve object detection, and the prediction box size setting information for each pixel in each feature layer is shown in Table 1. By obtaining lower spatial features and fusing them with high-level semantic features to generate a P2 detection layer, the model's ability to detect smaller objects is better improved [28].

III. EXPERIMENT

A. EXPERIMENTAL ENVIRONMENT

This article uses the Ubuntu20.04 system; the experimental environment is python3.8, pytorch1.10.0, and cuda11.3. All

TABLE 1. Information on the prediction box size setting for each pixel point in each feature layer.

Detection layer	Feature map	Receptive field	Anchor Boxes
P2	160 × 160 × 64	4 × 4	[7, 9, 9, 17, 17, 15, 13, 27]
P3	80 × 80 × 128	8 × 8	[21, 28, 36, 18, 23, 47, 35, 33]
P4	40 × 40 × 256	16 × 16	[58, 29, 43, 60, 82, 46, 66, 88]
P5	20 × 20 × 512	32 × 32	[133, 77, 111, 135, 206, 137, 197, 290]

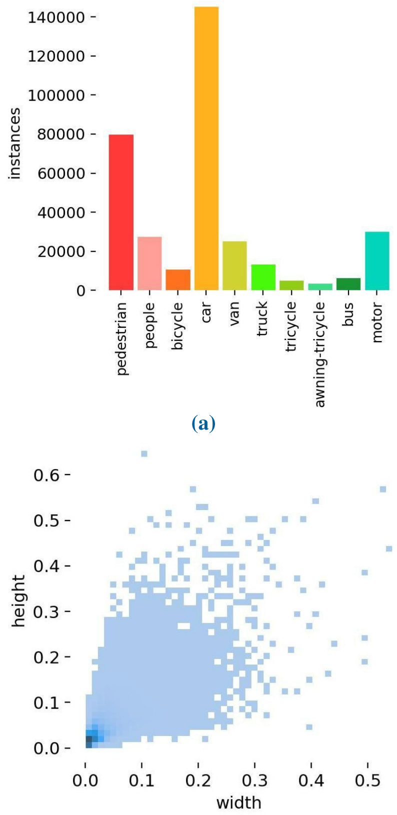
TABLE 2. Training parameters.

Training Parameters	Value(Category)
Epoch	200
Batch size	8
Image size	640 × 640
Selected model	YOLOv5x,YOLOv5 + P2,YOLOv5 + ESPP, YOLOv5 + SCEP,YOLOv5 + Ensemble
Model scaling factor	depth: 1.33 width: 1.25

models are run on the NVIDIA RTX3090 GPU, trained, validated and tested under the same hyperparameters. The specific parameters of the experiment are shown in Table 2.

B. INTRODUCTION TO THE DATASET

This paper selects the VisDrone2021 dataset to train and evaluate the model, which is gathered by the AISKYEYE team at Tianjin University's machine learning and data mining lab [29]. All benchmark datasets were taken by drones, including 288 video clips, 261908 frames, and 10209 static images, of which 6471 were selected as training sets, 3190 test sets, and 548 validation sets. There are ten categories of images with 2.6 million labels. Figure 7(a) is the number of labels for each category, and 7(b) is a distribution map of all label sizes in the training set. The horizontal and vertical coordinates in 7(b) represent the width and height of the label box, respectively. It can be noticed that the lower left corner has a larger concentration of points, indicating the presence of more small objects in the dataset, reflecting the general situation of UAVs in real-world application scenarios.



(a)

(b)

FIGURE 7. Results of the dataset used in this paper's attribute visualization.

C. EVALUATION CRITERIA

Currently, the most often used measures for assessing the performance of object detection algorithms are Precision, Recall, AP (average precision), mAP (mean AP), Params (number of parameters in the model), FLOPs (number of floating point operations), and FPS (frames per second of the image processed). In this paper, Precision, mAP0.5, mAP0.5:0.95, Params, FLOPs, and FPS are selected as the evaluation metrics of our model [30].

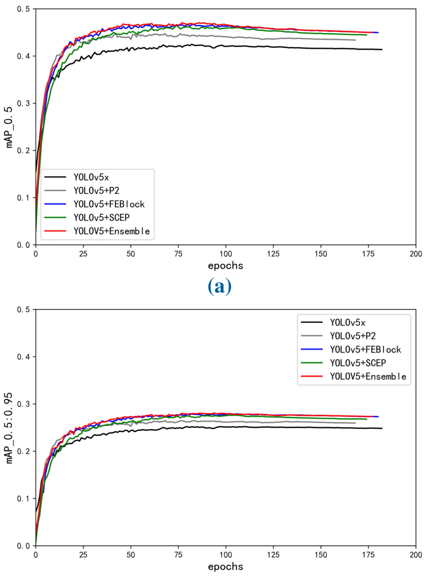
D. ABLATION EXPERIMENTS

To verify the effectiveness of the proposed feature enhancement block FEBlock, the self-feature expansion plate SCEP and the addition of a more shallow feature map as a detection layer. In this paper, ablation experiments are conducted to evaluate the impact of different modules on the performance of the UAV capture scene object detection algorithm under the same experimental conditions. YOLOv5x version 6.1 is chosen as the baseline model for the ablation experiments. The input image resolution was set to 640 × 640, and 200 epochs were trained. The experimental performance comparison between different models is shown in Table 3 and Figure 8.

It can be seen from Table 3 that the ablation experiment results B, mAP0.5 and mAP0.5:0.95 are 2.2% and 1.3% higher than the YOLOv5 baseline, respectively, indicating that adding a shallower feature map as a detection layer can improve the detection effect of small objects. Model C performs feature enhancement for each maximum pooling in SPP. Compared with Model A, mAP0.5 is increased by 4.2% and the accuracy is increased by 3.1%. The accuracy is the highest among all models, which indicates that embedding the FEBlock into the SPP is better than the original SPP. Model D splices SCEP and SPP separately, and SCEP expands and fuses the feature information. The precision is 1.6% higher than that of Model A, but 1.5% lower than that of Model B, which better proves that the embedding of FEBlock

TABLE 3. Performance comparison between different algorithms.

	Models	Precision	mAP0.5	mAP0.5 : 0.95	FPS	Params(M)	FLOPs(G)	Inference(ms)	Time(h)
A	YOLOv5x	56.4	42.5	25.2	43	86.2	203.9	19.9	11.1
B	YOLOv5 + P2	57.7	44.7	26.5	32	87.9	230.9	22.5	12.2
C	YOLOv5 + ESPP	59.5	46.7	27.9	25	99.7	234.9	32.2	16.7
D	YOLOv5 + SCEP	58.0	46.5	27.4	23	104.6	236.8	32.6	16.9
E	YOLOV5 + Ensemble	59.1	47.1	28.1	21	117.8	272.4	33.0	18.7



(a)

(b)

FIGURE 8. mAP0.5, mAP0.5:0.95 for the ablation experiment.

into SPP improves the detection accuracy of the model. Model E integrates all improvements together, mAP0.5 is 4.6% higher than the baseline, and mAP0.5:0.95 is 2.9% higher, which is the largest improvement among all models. These improvements together lead to a significant increase in the number of parameters and a significant negative correlation in computational complexity. Still, the accuracy and mAP have been greatly improved, which can improve the effect of UAV capture scene object detection.

E. MODEL SELECTION EXPERIMENTS

In this paper, we conduct comparative experiments on the VisDrone2021 dataset according to different input image resolution parameters (320, 480, 640, 1024, 1504). The experimental results are shown in Table 4. The image captured by UAV has a high resolution,and due to the low resolution of small objects, high-resolution images can retain more detailed features. It can be seen from Table 4 that when the training input resolution parameter of YOLOv5x is 1024, mAP0.5 is 11.9% higher than 640, and mAP0.5 of the improved YOLOv5x is 9.7% higher than 640, and FPS remains unchanged. It can be seen that improving the resolution does greatly improve the detection accuracy of the model, and because the network structure and scale have not changed, increasing the resolution of the input image does not affect real-time performance. For images with a resolution parameter of 1504, the amount of calculation increases significantly, and the growth of mAP0.5 is much slower than that of 1204. Therefore, it is necessary to note that when the input resolution is too high, the model calculation is too complex and prone to over-fitting, which leads to a decrease in detection accuracy. In contrast, when the resolution is reduced, mAP0.5 drops significantly, and FPS is unchanged; detecting small objects is unsuitable for reducing the resolution.

TABLE 4. Effect of different input image resolutions during training.

Input Resolution	Method	mAP0.5	FPS_{1024}
320 × 320	YOLOv5x	23.1	43
480 × 480		35.0	43
640 × 640		42.5	43
1024 × 1024		54.4	43
1504 × 1504		56.9	43
320 × 320	Improved YOLOv5x	30.2	21
480 × 480		40.9	21
640 × 640		47.1	21
1024 × 1024		56.8	21
1504 × 1504		58.2	21

Note: Input Resolution represents the resolution of the input image during training, FPS_{1024} represents the resolution of the detection image of 1024 × 1024.

In addition, the performance changes of the model under different detection image resolutions are shown in Table 5. Different resolutions of the input images will affect the speed and accuracy of the detection stage. With the decrease in the input's resolution during detection, mAP0.5 continues to decrease and FPS continues to increase. When the resolution of the input increases during detection, PFS decreases continuously and has a greater impact. In addition, with the increase of input image resolution, mAP0.5 increases rapidly and then decreases slowly, indicating that the resolution will affect the detection accuracy. When the resolution increases from 1024 to 1504, mAP0.5 begins to decline. The resolution of the input image during training differs too much from that during detection, which will make the characteristic parameters of the model not match the detection image, resulting in the detection accuracy not rising and falling. Based on the careful consideration of Table 4 and Table 5, the resolution parameter of 1024 is selected as the image resolution during training and detection. The detection accuracy is higher and can better meet the actual project requirements.

F. COMPARATIVE EXPERIMENTS

This study compares the method with the most recent YOLO family of object detection algorithms, primarily assessing the algorithm's detection accuracy and detection speed to

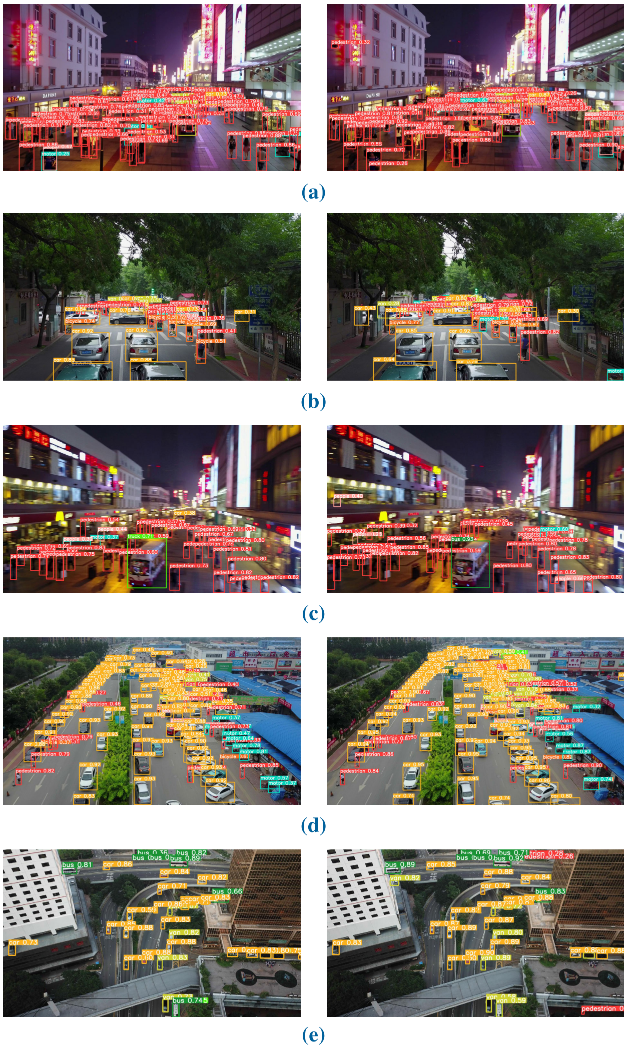
TABLE 5. Variation in model performance at different detection image resolutions.

Input Resolution	Method	mAP0.5	FPS
320 × 320	YOLOV5x	25.4	53
	Improved YOLOv5x	31.2	30
480 × 480	YOLOV5x	36.4	53
	Improved YOLOv5x	41.2	32
640 × 640	YOLOV5x	42.5	46
	Improved YOLOv5x	47.1	29
1024 × 1024	YOLOV5x	54.2	26
	Improved YOLOv5x	56.8	15
1504 × 1504	YOLOV5x	54.4	13
	Improved YOLOv5x	56.7	7

show the upgraded YOLOv5x algorithm's superiority over other algorithms. For a quantitative investigation of tiny target identification outcomes in UAV capture scenarios, the enhanced YOLOv5x algorithm is compared with the YOLOv5x, YOLOX [31], YOLOv6 [32], and YOLOv7 [33] algorithms, all selecting the largest network structure of each algorithm. The findings are displayed in Table 6. In addition, we compared some of the latest improved YOLOv5 algorithms. The network can mine more feature information by improving the attention mechanism and adding it to the backbone network. Most of the algorithms are equivalent to the method in this paper in accuracy, which further improves the detection accuracy of small targets, but the method in this paper has more advantages in real-time performance.

It can be seen from Table 6 that the detection accuracy of the YOLOv6 algorithm is relatively low. Because the parameters are low and can be deployed in some embedded devices, YOLOv6 has more advantages in a real production environment. Still, the benefits of the pure algorithm effect are not obvious. In contrast, although the detection accuracy of the YOLOX algorithm is slightly higher than that of the YOLOv6 algorithm, the number of parameters and calculations is significantly increased, and the real-time performance is greatly reduced. The detection effect of small objects is not ideal.

On the whole, the proposed algorithm introduces FEBlock in spatial pyramid pooling to enhance the features of feature maps at different scales and focus more on small object regions, improving the algorithm's small object feature extraction capability; at the same time, the SCEP module is spliced in after the enhanced spatial pyramid pooling to feature fusion further and expand feature information on the feature maps output from the enhanced spatial pyramid pooling, enhancing the feature representation capability and improving the network's effect on small object detection. Finally, a small object detection layer is added to enhance the network's detection effect on small objects. Although the improved algorithm in this paper is not ideal in terms of real-time performance, it enhances the feature extraction capability of small objects. It has a good detection effect on small objects in UAV capture scenarios.


(a)

(b)

(c)

(d)

(e)

FIGURE 9. Comparison chart of test results.

G. ALGORITHM VALIDITY ANALYSIS

To verify the object detection effect of the improved algorithm in the actual scene, this paper uses the representative images in the VisDrone2021 test set to test and make a visual comparison. The detection effect is shown in Figure 9. In Figure 9, the left side is the YOLOv5 baseline, and the right side is the improved algorithm. Select the night street, road occlusion, blur distortion, and high altitude scenes as the detection object. Figure 9(a) shows a real scene of a pedestrian street at night, with a large number of dense pedestrians and dim light; the improved model improves the detection effect in dim scenes. Some occlusions and overlaps exist in Figure 9(b) and 9(d). The improved model can detect small occluded objects and distinguish different types. Figure 9(c) in the presence of blurred image distortion, the model can still detect the fuzzy object stability, which can be an excellent response to the actual situation. Figure 9(e) is a picture taken at a high altitude. The vehicle on the road is very small, but it can still improve the detection effect, indicating that the model can detect small objects [34].

TABLE 6. Comparison experiments of the latest YOLO series target detection algorithms.

Models	mAP0.5	mAP0.5 : 0.95	Params(M)	FLOPs(G)	FPS	Inference(ms)
YOLOv5	42.5	25.2	86.2	203.9	43	19.9
YOLOX	40.8	24.2	99.1	281.9	26	26.5
YOLOv6	38.4	22.8	58.5	143.8	45	22.3
YOLOv7	51.3	30.0	70.8	188.2	59	14.9
Ours	56.8	35.4	117.8	272.4	21	33.0

The improved algorithm increases the receptive field through feature enhancement blocks, generates adaptive weights for different receptive fields, and improves the ability of the model to extract small objects at different scales. Feature enhancement blocks are fused into SPP, and feature enhancement is performed for the results of each maximum pooling to generate new features containing multi-scale contextual information by weighting the fused contextual features. The feature enhancement blocks are introduced in SPP to improve the global feature extraction ability by weakening the background noise interference. After splicing with SCEP, the feature information is further fused and expanded, so that the model has better robustness to cope with the actual situation. The improved algorithm in Figure 9 reduced false detection and missed detections when dealing with small dense objects and reduced the impact of environmental illumination changes. Moreover, it can detect pedestrians occluded by trees and distinguish distant pedestrians from vehicles, and the detection effect of distorted images in blurred scenes is still improved. Overall, the detection accuracy of the improved algorithm has been improved to a certain extent, enhancing the detection effect of small dense objects.

IV. CONCLUSION AND FUTURE WORK

This paper proposes an object detection algorithm based on improved YOLOv5 for UAV capture scenes, which is intended to improve the detection accuracy of small-size objects and small dense objects.

Firstly, a feature-enhanced block, FEBlock, is proposed, which integrates information without deepening the network structure by combining different channel information. Adaptive weights are generated for different receptive field features by convolution. The main weights are assigned to shallow feature maps to focus more attention on dense small object regions and improve small object feature extraction. Then FEBlock is integrated into SPP to generate enhanced spatial pyramid pooling ESPP, and feature enhancement is performed for each maximum pooling result. By weighted fusion of context features, new features containing multi-scale context information are generated to weaken the interference of background noise and have better feature representation ability. In addition, the self-characteristic expansion plate SCEP is proposed and stitched with ESPP to generate a new spatial pyramidal pooling ESPP-S, which further improves the feature extraction capability of the network by achieving the fusion and expansion of feature information through compression, non-linear mapping, and expansion with its own module. Finally, based on the large, medium, and small detection layers of the YOLOv5 network model, a shallower feature map is added as the detection layer according to the characteristics of the data set, which improves the detection ability of the model for smaller objects. The final experimental results show that the precision of the improved YOLOv5 algorithm reaches 59.1%, mAP0.5 can reach 47.1%, and mAP0.5:0.95 reaches 28.1%, which is a good improvement compared with YOLOv5x. And the training model mAP0.5 can reach 56.8% under the input resolution of 1024 × 1024, which has a good effect on the actual detection. In this paper, the detection accuracy has been significantly improved, but the parameter quantity and calculation amount have been greatly improved, the real-time performance has been greatly reduced, and the UAV airborne computing resources are limited. Therefore, how the lightweight and efficient model can be further studied.

REFERENCES

[1] B. Jiang, R. Qu, Y. Li, and C. Li, "Object detection in UAV imagery based on deep learning: Review," *Acta Aeronautica et As-Tronautica Sinica*, vol. 42, no. 4, 2021, Art. no. 524519.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[3] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[5] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2016, pp. 21–37.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[13] Z. Q. Feng, Z. J. Xie, Z. W. Bao, and K. W. Chen, "Real-time dense small object detection algorithm for UAV based on improved YOLOv5," (in Chinese), *Acta Aeronautica et Astronautica Sinica*, vol. 44, no. 3, p. 327106, 2023, doi: 10.7527/S1000-6893.2022.27106.

[14] 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.

[15] M. Hong, S. Li, Y. Yang, F. Zhu, Q. Zhao, and L. Lu, "SSPNet: Scale selection pyramid network for tiny person detection from UAV images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] J. Zhao, X. Zhang, J. Yan, X. Qiu, X. Yao, Y. Tian, Y. Zhu, and W. Cao, "A wheat spike detection method in UAV images based on improved YOLOv5," *Remote Sens.*, vol. 13, no. 16, p. 3095, Aug. 2021.

[17] W. Liu, K. Quijano, and M. M. Crawford, "YOLOv5-tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8085–8094, 2022.

[18] X. Zhang, Y. Feng, S. Zhang, N. Wang, and S. Mei, "Finding nonrigid tiny person with densely cropped and local attention object detector networks in low-altitude aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4371–4385, 2022.

[19] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Jun. 2015.

[20] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[21] H. Liu, F. Sun, J. Gu, and L. Deng, "SF-YOLOv5: A lightweight small object detection algorithm based on improved feature fusion mode," *Sensors*, vol. 22, no. 15, p. 5817, Aug. 2022.

[22] J. Wang, T. Xiao, Q. Gu, and Q. Chen, "YOLOv5_CSL_F: YOLOv5's loss improvement and attention mechanism application for remote sensing image object detection," in *Proc. Int. Conf. Wireless Commun. Smart Grid (ICWCSG)*, Aug. 2021, pp. 197–203.

[23] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 13–19.

[24] T. Deng, X. Liu, and L. Wang, "Occluded vehicle detection via multi-scale hybrid attention mechanism in the road scene," *Electronics*, vol. 11, no. 17, p. 2709, Aug. 2022.

[25] J.-L. Zhang, W.-H. Su, H.-Y. Zhang, and Y. Peng, "SE-YOLOv5x: An optimized model based on transfer learning and visual attention mechanism for identifying and localizing weeds and vegetables," *Agronomy*, vol. 12, no. 9, p. 2061, Aug. 2022.

[26] L. Wang, Y. Cao, S. Wang, X. Song, S. Zhang, J. Zhang, and J. Niu, "Investigation into recognition algorithm of helmet violation based on YOLOv5-CBAM-DCN," *IEEE Access*, vol. 10, pp. 60622–60632, 2022.

[27] S. Luo, J. Yu, Y. Xi, and X. Liao, "Aircraft target detection in remote sensing images based on improved YOLOv5," *IEEE Access*, vol. 10, pp. 5184–5192, 2022.

[28] X. Luo, Y. Wu, and F. Wang, "Target detection method of UAV aerial imagery based on improved YOLOv5," *Remote Sens.*, vol. 14, no. 19, p. 5063, Oct. 2022.

[29] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," 2018, *arXiv:1804.07437*.

[30] H. Lv, H. Yan, K. Liu, Z. Zhou, and J. Jing, "YOLOv5-AC: Attention mechanism-based lightweight YOLOv5 for track pedestrian detection," *Sensors*, vol. 22, no. 15, p. 5903, Aug. 2022.

[31] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding Yolo series in 2021," 2021, *arXiv:2107.08430*.

[32] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[33] 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*.

[34] H. Gong, T. Mu, Q. Li, H. Dai, C. Li, Z. He, W. Wang, F. Han, A. Tuniyazi, and H. Li, "Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images," *Remote Sens.*, vol. 14, no. 12, p. 2861, Jun. 2022.

ZHEN LIU received the B.S. degree in electrical engineering and automation from the Nanhang Jincheng College, Nanjing, China, in 2021. He is currently pursuing the M.S. degree with the College of Intelligent Equipment, Shandong University of Science and Technology.

His current research interests include deep learning, object detection, and intelligent control.

XUEHUI GAO (Member, IEEE) received the M.S. degree in detection technology and automatic equipment from the School of Information and Electrification Engineering, Shandong University of Science and Technology, Qingdao, China, in 2006, and the Ph.D. degree in control science and engineering from the Beijing Institute of Technology, Beijing, China, in 2017.

Since 2018, he has been an Assistant Professor with the School of Intelligent Equipment, Shandong University of Science and Technology, Tai'an, China. His current research interests include hysteresis nonlinear system adaptive control, system identification, and neural networks.

YU WAN was born in Shandong, China, in 1998. He received the bachelor's degree from the Qingdao University of Technology. He is currently pursuing the master's degree with the Shandong University of Science and Technology.

His recent research interests include image processing and evolutionary algorithms.

JIANHAO WANG was born in Hebei, China, in 1997. He received the bachelor's degree from the Hebei University of Science and Technology. He is currently pursuing the master's degree with the Shandong University of Science and Technology.

His recent research interests include image processing and machine vision.

HAO LYU was born in Shandong, China, in 1998. She received the bachelor's degree from the Shandong University of Science and Technology, where she is currently pursuing the master's degree.

Her recent research interest includes adaptive control.