An Efficient Method for Detecting Dense and Small Objects in UAV Images

Chenyang Li, *Student Member, IEEE*, Suiping Zhou, Hang Yu, *Member, IEEE*, Tianxiang Guo, Yuru Guo, and Jichen Gao

***Abstract*—Object detection in unmanned aerial vehicle (UAV) images is an important and challenging task for many applications, which often needs highly efficient detection algorithms to meet the accuracy and real-time requirements of the applications. In this article, we investigate efficient mechanisms for detecting dense and small objects in UAV images. Specifically, 1) kernel K-means is used to obtain optimal anchors for dense and small object detection; 2) a spatial information enhancement module is proposed to improve the detection accuracy of dense objects by extracting object spatial location information; 3) a Coord_C3 module is proposed to improve the receptive field of the network and to reduce the number of network parameters; and 4) a small detection head is added in the Head of the network and skip connections are employed in the Neck of the network to improve the detection accuracy of small objects. Experimental results on the VisDrone-2019, LEVIR-ship, and Stanford Drone datasets show that our method not only has higher detection accuracy but also runs faster compared to state-of-the-art detection methods.**

***Index Terms*—Deep learning, kernel K-means, object detection, spatial information, unmanned aerial vehicle (UAVs).**

I. Introduction

With the development of the unmanned aerial vehicle (UAV) technology, object detection in UAV images has a wide range of applications, including urban environment monitoring [1], land utilization planning [2], forest fire monitoring [3], traffic management [4], and military [5]. Images captured under a UAV's field of view are more complex than images of natural scenes. Specifically, 1) UAV images are variable and complex, and object distribution may be dense or sparse. Existing object methods often demonstrate low robustness in this context and 2) as the UAVs are usually far from the ground, the objects in the captured images may be small, which makes it hard to extract the real contours of the objects with the existing object detection methods. Thus, detecting these dense and small objects in UAV images is an important and challenging task.

Manuscript received 12 November 2023; revised 30 January 2024; accepted 1 March 2024. Date of publication 5 March 2024; date of current version 19 March 2024. This work was supported in part by the One Hundred Person Project of the Shaanxi under Grant 10253180002, in part by Hainan Province Science and Technology Special Fund under Grant ZDKJ2021019, and in part by the Research on Experimental Simulation and Processing Methods for SAR Load System Data under Grant 99903220109. *(Corresponding authors: Suiping Zhou; Hang Yu.)*

The authors are with the School of Aerospace Science and Technology, Xidian University, Xi'an 710126, China (e-mail: chenyangli@stu.xidian.edu.cn; spzhou@xidian.edu.cn; yuhang9551@163.com; 21131110570@stu.xidian.edu.cn; 21131213395@stu.xidian.edu.cn; gjchen1999@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3373231

Many methods have been proposed for detecting small objects in UAV images [6], [7], [8], [9], [10], [11], [12], [13], [14]. The authors in [8] enhanced the accuracy of small object detection through alignment fusion of shallow spatial features and deep semantic features, employing candidate region feature alignment. Liu et al. [11] improved the detection accuracy of small object in UAV images by connecting two ResNet units of equal width and height based on YOLOv3. Two data enhancement strategies and distance metrics are proposed in [12] for improving detection accuracy of small objects. Zhang et al. [13] proposed a spatial logical aggregation network (SLA-NET) with morphological transformations, which enables the extraction of fine-grained features of small objects through multiple plug-and-play dynamic fusion modules. The authors in [14] proposed a multibranch parallel feature pyramid network (MPFPN) and focuses attention on object information, which improves the detection accuracy of small objects. However, the computational cost of these methods is generally large and can hardly meet the real-time requirements of many UAV applications. In addition, the detection accuracy of these methods may be decreased when the objects are densely distributed.

Some methods have been proposed for detecting dense objects in UAV images. Xu et al. [15] proposed an advanced foreground-enhanced attention Swin transformer (FEA-Swin) framework that integrates contextual information into the original backbone of the Swin transformer. To avoid losing the information of small objects, an improved weighted bidirectional feature pyramid network (BiFPN) is proposed. To balance the detection accuracy and efficiency, an efficient bidirectional feature pyramid network neck is introduced. A novel semantic embedding density adaptive network (SDANet) was proposed in [16], which designs a new density matching algorithm to obtain each object by partitioning the clustering proposal and performing hierarchical and recursive matching of the corresponding centers. Ye et al. [17] proposed a backbone network utilizing involution and self-attention, capable of extracting effective features from complex objects. Furthermore, they introduced a multiscale feature fusion module to address the issue of large number of small objects in UAV images through multiscale object detection and feature fusion. However, due to the large number of parameters involved, these methods are difficult to meet the real-time requirements of many UAV applications.

In addition, many methods have been proposed for real-time object detection in UAV images [18], [19], [20], [21], [22]. Zhang et al. [18] achieved real-time object detection

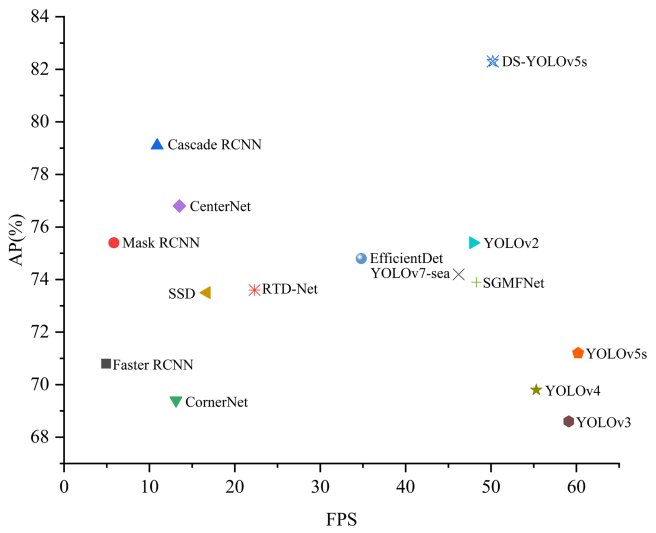

Fig. 1. FPS and AP comparison of object detection methods on the VisDrone-2019.

implementation for UAVs by introducing channel-level sparsity in the convolutional layer. This was accomplished through the application of L1 regularization to the channel scaling factor and the removal of less informative feature channels using clipping techniques. The authors in [21] achieved real-time object detection by replacing standard convolution with depth-separable convolution based on YOLOv3. A convolutional multihead self-attention (CMHSA) based on efficient convolutional transformer block (ECTB) was proposed in [22] for achieving real-time object detection. CMHSA employs a convolutional projection instead of a positional-linear projection, which reduces computational cost. These methods help to reduce the computational cost for object detection in UAV images, they often results in low detection accuracy, particularly for dense and small objects.

To address these problems, we investigate efficient mechanisms (DS-YOLOv5s) for detecting dense and small objects in UAV images with high accuracy and low computational cost. Fig. 1 summarizes the performance of our method as compared with the state-of-the-art methods.

The major contributions of our work are as follows.

1) A kernel K-means is used to obtain optimal anchors for dense and small object detection.
2) A spatial information enhancement (SIE) module is proposed to improve of detection accuracy of dense objects by extracting object spatial location information.
3) A Coord_C3 module is proposed to improve the receptive field of the network and to reduce the number of network parameters.
4) A small detection head is added in the Head of the network and skip connections are employed in the Neck of the network to improve the detection accuracy of small objects.

The rest of this article is organized as follows: Section II describes the related work. Section III provides a detailed description of the proposed method for dense and small object detection. The experimental results are presented and analyzed in Section IV. Further analysis is discussed in Section V. Finally, Section VI concludes this article.

II. Related Work

As our method is based on YOLOv5s, in this section, we will first introduce the principles of YOLOv5s, then describe some related work in feature extraction for object detection in UAV images.

A. YOLOv5s

YOLOv5 [23] is the fifth generation of You Only Look Once (YOLO) [24], a state-of-the-art object detection network. YOLOv5 includes YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, which are classified mainly according to their size and computational complexity. Compared to other networks, YOLOv5s has the advantages of fewer parameters and faster detection speed. The structure of YOLOv5s includes the Input, Backbone, Neck, and Head, as shown in Fig. 2.

Input: During the initial stage of processing images, the input image is adjusted to conform to the size of the model's input through normalization and adaptive scaling. The Mosaic data augmentation [25] and adaptive anchor [26] methods are used to improve the inference speed of the network, and enhance the robustness of the network.

Backbone: The Backbone of YOLOv5s is the improved CSP-Darknet53 network [27], which combines of CBS, C3, and SPPF [28] modules for refined feature information extraction. CSPDarknet53 effectively enhances the learning capability of convolutional neural networks (CNNs) while simultaneously reducing computational cost.

Neck: It serves the purpose of connecting the Backbone network with the prediction head network, facilitating the acquisition and transmission of feature information. It consists of two networks. The feature pyramid network (FPN) [29] has an up–down structure that upsamples and fuses the underlying feature information to obtain the predicted feature map, while the path aggregation network (PANet) [30] uses a down–up structure to fuse the FPN feature map to complement the FPN structure .

Head: The YOLOv5s contains three object detection heads which correspond to three different sizes of feature maps. Each grid on the feature map is predefined with three anchors of different aspect ratios, which is used to store anchor-based position and classification information in the feature map channel dimension for object prediction and regression. The prediction frame is calibrated by CIoU loss [31], and the optimal prediction frame is obtained by nonmaximum suppression (NMS) [32].

In this study, YOLOv5s will serve as the basis as well as the benchmark. We made various enhancement to YOLOv5s to address the accuracy and real-time requirements for detecting dense and small objects in UAV images.

B. Feature Extraction for Object Detection

Traditional convolutional networks use an up–down substructure, where the expressiveness of the object's shallow features

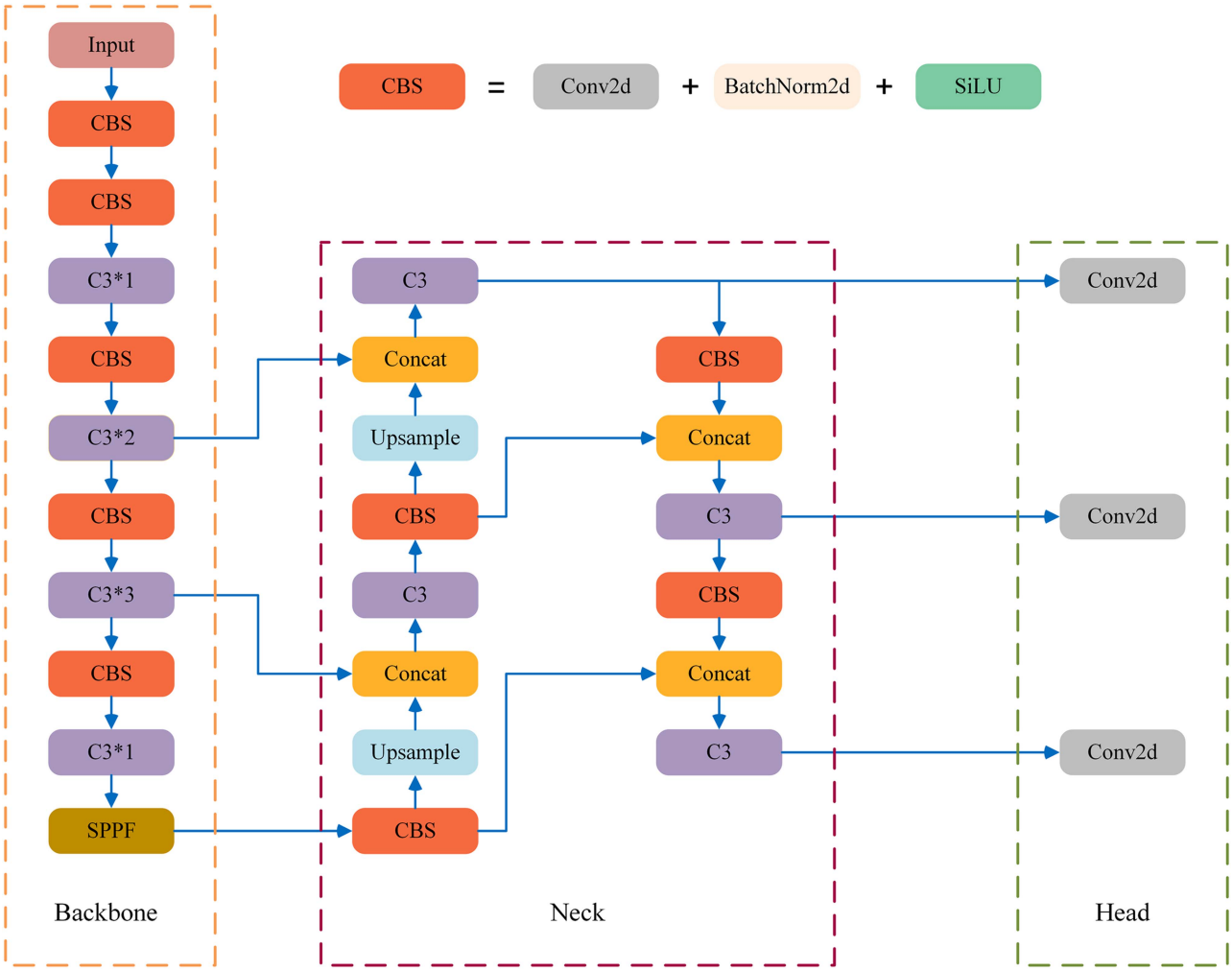

Fig. 2. Structure of YOLOv5s network.

decreases as the convolutional layers become deeper. As the semantic information of smaller objects often appears in shallower feature mappings, deeper convolutional layers may result in the loss or complete disappearance of object feature information. Therefore, common methods for solving this problem include feature fusion, receptive field enhancement, and anchor matching.

Feature fusion is widely used in object detection. Jin et al. [33] combined the semantic information of feature context at different scales with an expansive convolution method. Han et al. [34] used deconvolution to enhance the feature representation of ship objects. Wang et al. [35] designed SSS-YOLO to fuse feature and semantic information by a path-enhanced fusion network. A two-way convolution network (TWCNet) is proposed in [36] to process both shallow and deep feature information. The authors in [37] proposed a graph feature-enhanced selective assignment network (GSANet) that uses graph convolutional networks to obtain topological information between ground objects to enhance representational features. For receptive field enhancement, Zhao et al. [38] grew candidate regions from multiple receptive fields and combined the contextual information of the candidate frames to improve the detection accuracy of the object. Dai et al. [39] fused down–up and up–down feature maps to enhance the receptive fields of small object features. To expand the receptive field of the convolution kernel, Wang et al. [40] introduced large kernel convolution by replacing the small convolution kernel with two parallel rectangular convolution kernels. Expanding the receptive field while maintaining the ability to capture local detailed features improves the object detection accuracy. As to anchor matching, Fu et al. [41] adopted an anchor-free strategy to detect small ships in SAR images. Xu et al. [42] used an improved K-means++ algorithm to optimize the anchors and to alleviate the difficulty in optimizing multiscale features of ships. Liang et al. [43] proposed a concise analytical geometry algorithm to calculate ship orientation and gradually refine the keypoints to establish an accurate orientated bounding box.

III. DS-YOLOv5s

Due to the fact that UAVs fly at high altitudes, this results in a high proportion of small objects in the image, which are densely distributed. In addition, it is often difficult to balance between the high computational demand and the limited arithmetic power

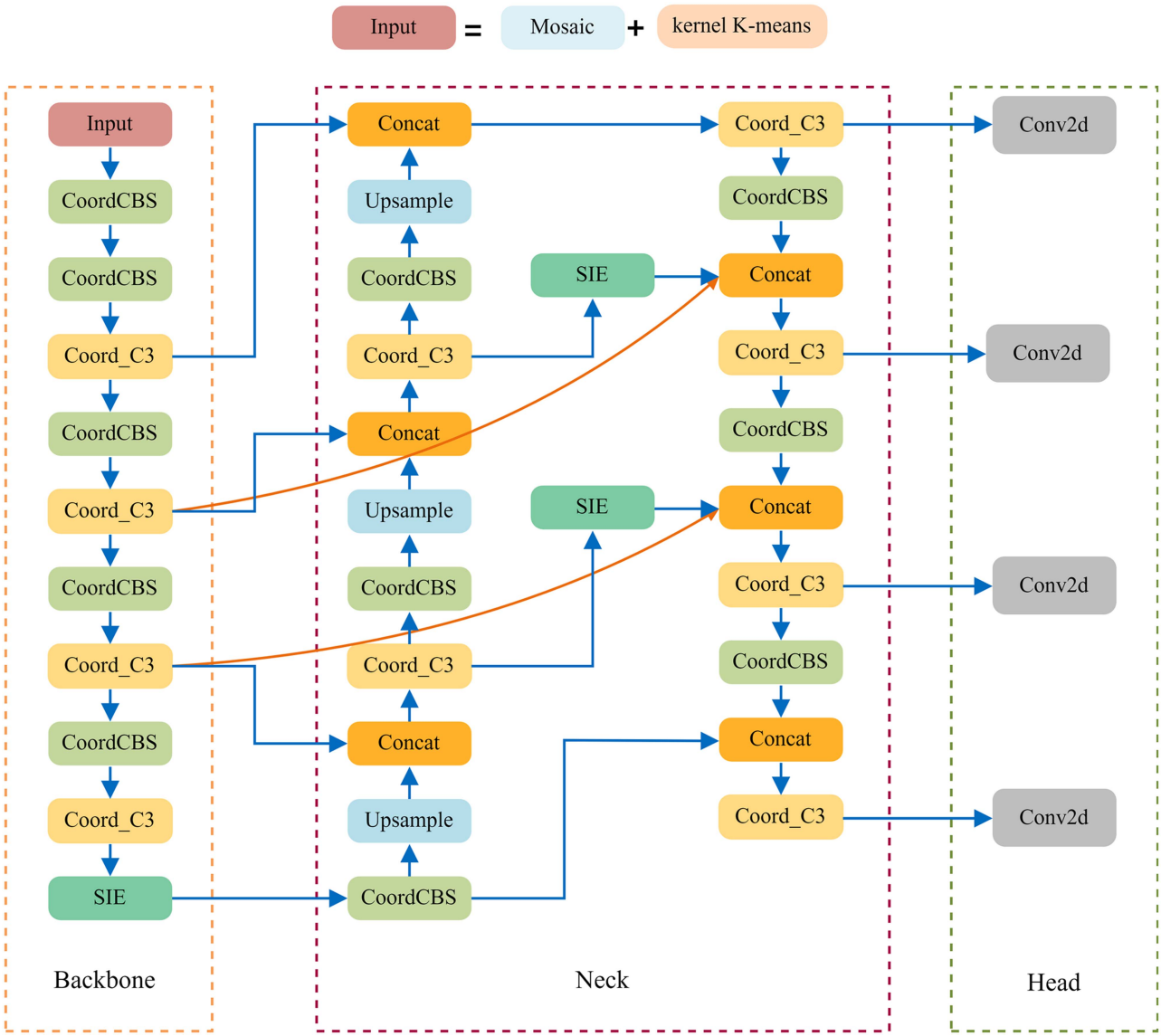

Fig. 3. Structure of DS-YOLOv5s network.

of low-power chips of UAVs. To address these problems, we propose a method for detecting dense and small objects in UAV images based on YOLOv5s. First, the anchor of the dataset is optimized by kernel K-means clustering algorithm. Then, we introduce an SIE module in the Backbone and Neck of the network, which enhances the location information of dense objects in the network by extracting spatial feature information. At the Backbone and Neck network, based on CoordConv [44], the Coord_C3 module is proposed to replace the C3 of YOLOv5s, which can improve the receptive field of the network and reduce the number of parameters of the network. In addition, at the Head of the network, a detection head for a small object is added to improve the detection accuracy of small objects. Finally, skip connections are introduced to fuse the shallow and deep features, which can improve the feature sensing ability of the network and further improve detection accuracy of dense and small objects. The DS-YOLOv5s network structure is shown in Fig. 3.

A. *Anchor Optimization*

YOLOv5s defines three initial anchors, and optimizes the anchor using the K-means clustering algorithm [45]. However, K-means requires manual initialization of the clustering center, which results in low clustering accuracy. Existing algorithms use K-means++ [46] to avoid manual initialization of the cluster centers, but result in higher computational complexity.

Dhillon et al. [47] proposed kernel K-means based on K-means. Unlike traditional K-means, kernel K-means uses kernel functions to map data into high-dimensional space before clustering. This algorithm can effectively process a nonlinear distribution datasets and improve clustering accuracy.

Ensuring reasonable anchor is a crucial requirement for improving detection accuracy of objects. The UAV object detection dataset usually contains multiple categories of objects, each of which has a different size and with a nonlinear

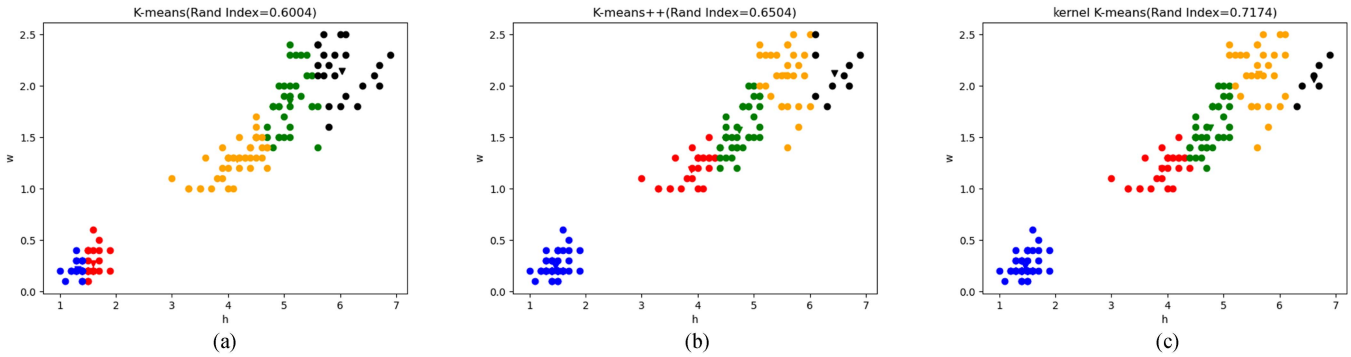

Fig. 4. Comparison of clustering results. (a) K-means (Rand Index=0.6004). (b) K-means++ (Rand Index=0.6504). (c) kernel K-means (Rand Index=0.7174).

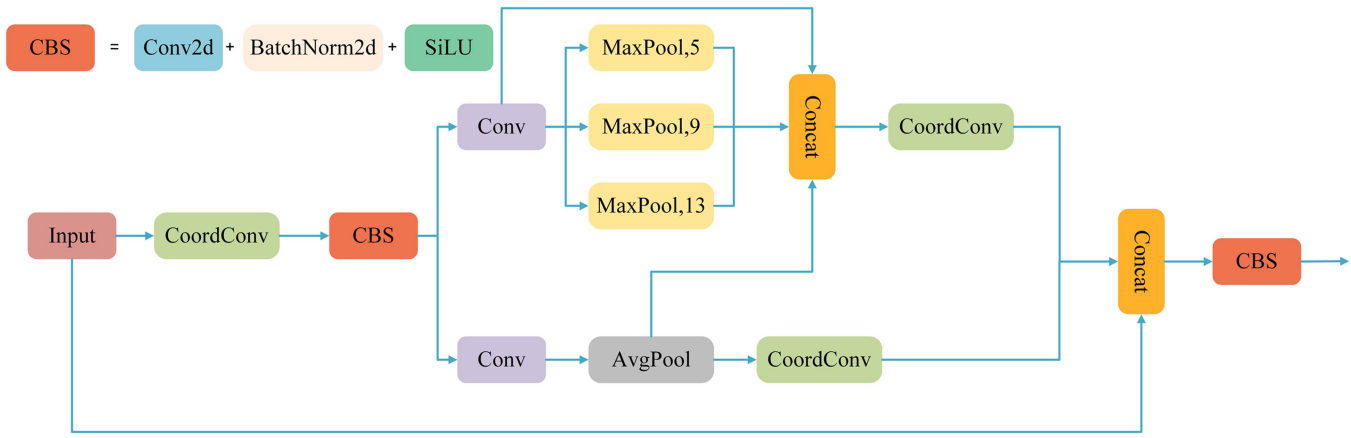

Fig. 5. Structure of the SIE moudle.

distribution. To obtain optimal anchors, this study introduces kernel K-means clustering at the input of the DS-YOLOv5s network to automatically find the reasonable anchor for object detection. Fig. 4 illustrates an example where the clustering center is set to 5 and the number of iteration is set to 10. Conduct experiments on simulated data, with the Rand index serving as the evaluation metric. The clustering process using K-means, K-means++, and kernel K-means took 0.18, 0.21, and 0.12 s, respectively. Notably, kernel K-means achieved the highest Rand index, indicating superior performance in terms of speed and accuracy. Consequently, this study utilizes kernel K-means for anchor frame optimization. To the best of our knowledge, this is the first application of kernel K-means in anchor optimization.

B. SIE Module

An SIE module is proposed to extract spatial location features in UAV images, weighting different channel features and spatial locations, and enhance the network's perception and positioning ability of object categories and location distributions. The SIE module structure is shown in Fig. 5. The location information of the feature maps is extracted through CoordConv [44], and the feature expression capability of the network is enhanced through the CoordCBS module. To enhance computational efficiency and improve feature representation, the feature maps is compressed using 1×1 convolution (Conv) in the channel dimension. This process effectively removes redundant channel features and reduces the number of parameters. Subsequently, the feature maps are concatenated in both height and width directions employing maximum pooling and global average pooling. The SIE module is expressed as the following equation:

$$\mathrm{SIE} = \mathrm{CBS}(I_5) \tag{1}$$

$$I_5 = ca(I_3, I_4, i) \tag{2}$$

$$I_4 = cc(I_2) \tag{3}$$

$$I_3 = cc(AP(c(I_1))) \tag{4}$$

$$I_2 = ca(MP_{5,9,13}(c(I_1)), c(I_1), AP(c(I_1))) \tag{5}$$

$$I_1 = \mathrm{CBS}(cc(i)) \tag{6}$$

where c, cc, ca, AP, MP, and i denote Conv, CoordConv, Concat, AvgPool, MaxPool, and Input, respectively.

C. Coord_C3 Module

To improve receptive field and reduce the number of parameters of the network, we proposed the Coord_C3 module with spatial information based on CoordConv, as shown in Fig. 6.

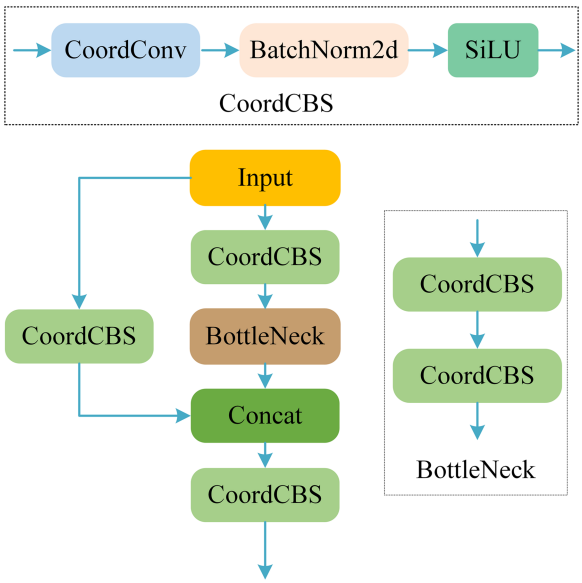

Fig. 6. Structure of the Coord_C3 moudle.

The Coord_C3 module includes the CoordCBS and BottleNeck modules. Through the multiscale exchange of shallow high-resolution and deep high-level semantic features, enhanced features containing dense small-object details and spatial information can be obtained. BottleNeck consists of two CoordCBS modules. It obtains spatial location information of the object while reducing the feature dimension and the number of layers in the feature map, thus reducing the computational complexity. The CoordCBS module consists of a CoordConv module, batch normalization (BatchNorm2d), and SiLU activation function, which can enhance the feature expression ability of CNNs while obtaining feature spatial information. The Coord_C3 expression is as follows:

$$\text{Coord\_}C3 = \text{CB}(Cc(\text{BN}(\text{CB}(i), \text{CB}(i)))) \tag{7}$$

$$\text{BN} = \text{CB}(\text{CB}(i)) \tag{8}$$

$$\text{CB} = \text{SiLU}(\text{BN2}(\text{CoordConv}(i))) \tag{9}$$

where CB, BN, Cc, BN2, and i represent CoordCBS, BottleNeck, Concat, BatchNorm2d, and Input, respectively.

D. Loss Function

EfficiCLoss is a loss function utilized in object detection, which combines the advantages of class balanced importance sampling (CBIS) loss and focal loss [48]. Its purpose is to improve the detection accuracy of small objects and expedite the convergence of the network. Compared to DIoU [31] and SIoU [49], EfficiCLoss exhibits significant effectiveness in enhancing the accuracy of model detection for small objects, and accelerating training. Consequently, we employ EfficiCLoss as the loss function of our network. The equations of EfficiCLoss are as follows:

$$L_{\text{effici}} = \alpha \cdot L_{\text{cbis}} + (1-\alpha) \cdot L_{\text{focal}} \tag{10}$$

TABLE I
EXPERIMENTAL CONFIGURATION

Platform	Name
The Operating system	Windows 11
CPU	I5-10400F
GPU	Nvidia RTX 1660S
Python	3.7
PyTorch	0.12.0
S Software	Pycharm 2022

$$L_{\text{cbis}} = \alpha \cdot L_{ce} + (1-\alpha) \cdot L_{\text{focal}} \tag{11}$$

$$L_{\text{focal}} = -\alpha(1-p_t)^n \log(p_t) + \alpha(p_t)^n \log(1-p_t) \tag{12}$$

where L_{effici} is the EfficiCLoss loss, L_{cbis} is the CBIS loss, L_{focal} is the focal loss, α is a balance coefficient, L_{ce} is CE loss, p_t is the predicted probability, and n is a balance factor. In this article, α and n are 0.6 and 0.5, respectively.

IV. Experiments and Analysis

To evaluate the performance of the proposed method, this study uses the VisDrone-2019 [50], LEVIR-ship [51], and Stanford Drone [52] datasets and compares the proposed method with existing object detection methods. All experiments are conducted with the same hardware and software environment, whose configurations are shown in Table I. Our network is implemented using PyTorch on a Windows 11 operating system, with an Nvidia GeForce RTX 1660S GPU and CUDA 11.3. The stochastic gradient descent [53] is employed as the optimizer, with an input image size of 512×512 pixels. The network is trained for 300 epochs, including 50 epochs of freeze training, with a batch size of 8. Subsequently, unfreeze training is conducted with a batch size of 4. To enhance the training process, a Mosaic [25] data augmentation strategy is employed. The initial learning rate is set to 0.01, with a minimum learning rate of 0.0001, which is adaptively adjusted based on the dataset characteristics. The momentum parameter and weight decay are set to 0.937 and 0.0005, respectively. During the testing stage, a postprocessing step utilizing NMS is applied.

A. Datasets

1) VisDrone-2019: The VisDrone2019 dataset was compiled by the AISKYEYE team at the Lab of Machine Learning and Data Mining, Tianjin University, China. This benchmark dataset comprises 288 video clips, with 261 908 frames and 10 209 static images. These recordings were captured using various drone-mounted cameras, offering a comprehensive representation of different aspects, including location (spanning 14 different cities across China, separated by thousands of kilometers), environment (urban and rural settings), objects (pedestrians, vehicles, bicycles, etc.), and scene density (ranging from sparse to crowded scenes). It is important to note that the dataset was collected using diverse drone platforms, with varying models,

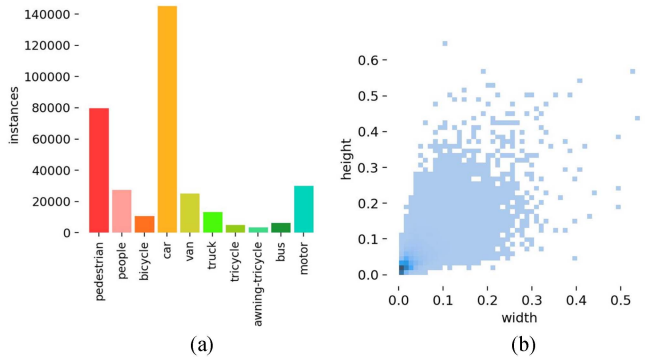

Fig. 7. Distribution of objects in VisDrone-2019. (a) Distribution of dataset categories. (b) Distribution of object sizes in the dataset.

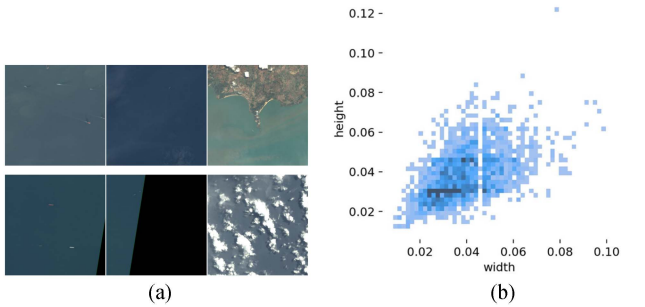

Fig. 8. Distribution of objects in LEVIR-ship dataset. (a) Images of LEVIR-ship dataset. (b) Distribution of object sizes in the dataset.

across different scenarios, and under various weather and lighting conditions. Manual annotation was carried out on these frames, resulting in over 2.6 million bounding boxes encompassing objects of interest, such as pedestrians, cars, bicycles, and tricycles. The distribution of features within the dataset is shown in Fig. 7. We employed this dataset to evaluate a method's capability in detecting dense objects.

2) Levir-Ship: The LEVIR-ship dataset comprises images captured by multispectral cameras on the Gaofen-1 and Gaofen-6 satellites. These images have a spatial resolution of 16 m and utilize only the R, G, and B bands. A total of 85 scenes were collected, with pixel resolutions ranging from 10 000×10 000 to 50 000×20 000. The original images were cropped to generate 1973 positive samples and 1923 negative samples, all of size 512×512 pixels. Detecting ships within this dataset poses challenges due to their relatively small size compared to the vast background. LEVIR-ship is a widely used dataset for object detection in remote sensing images, and its data distribution closely resembles that of UAV images. The distribution of features within the dataset is shown in Fig. 8. We utilized this dataset to evaluate a method's capability in detecting small objects.

3) Stanford Drone: The Stanford Drone dataset is an outdoor UAV dataset collected by the Computational Vision and Geometry Lab in the Department of Computer Science at Stanford University containing images and videos of various types of targets (not only pedestrians but also bicycles, skateboards, cars, buses, and golf carts). The dataset collects trajectory interaction

Fig. 9. Images of the Stanford Drone.

information of 20 k objects in eight different scenes using UAVs in an overhead view during crowded time periods on campus, and each object track is labeled with a unique ID suitable for object trajectory prediction and multiobject tracking. The number of videos in each scene and the percentage of each agent in each scene are shown in Fig. 9. We utilized this dataset to evaluate a method's capability in detecting dense and small objects.

B. State-of-the-Art Methods

We chose 14 object detect methods, e.g., Faster RCNN [54], Mask RCNN [55], Cascade RCNN [56], CornerNet [57], CenterNet [58], SSD [59], YOLOv2 [60], YOLOv3 [61], YOLOv4 [62], YOLOv5s [23], EfficientDet [63], SGMFNet [64], YOLOv7-sea [65], and RTD-Net [22], respectively, to validate the performance of the proposed method in object detect tasks in UAV images.

C. Evaluation Metrics

As primary accuracy evaluation metrics, standard average precision (AP), mean average precision (mAP), precision (P), and recall (R) are widely used in object detection for both natural and remote sensing images. To assess a model's efficiency, we measure frames per second (FPS), parameters (Params), and floating-point operations (FLOPs).

In the context of a model or classifier, the metric P represents the ability to accurately predict positive samples, with a higher value indicating superior performance. On the other hand, R represents the proportion of predicted positive samples relative to the total number of samples, and its performance aligns with that of P. It is worth noting that P and R have a mutual influence on each other. Generally, when P is high, R tends to be low, and vice versa. The metrics P, and R are computed using the following equations:

$$P = \frac{\text{TP}}{\text{TP+FP}} \tag{13}$$

$$R = \frac{\text{TP}}{\text{TP+FN}} \tag{14}$$

TABLE II
ABLATION EXPERIMENTS ON VISDRONE-2019 (%)

YOLOv5s [23]	SIE	Coord_C3	EfficiCLOSS	P	R	mAP@0.5	mAP@0.5:0.95
✓				71.2	63	75.8	27.5
✓	✓			71.9	64.5	76.1	29.3
✓		✓		72.1	64.9	76.4	29.5
✓			✓	71.6	65.1	76.3	30.3
✓	✓	✓		72.8	66.2	77.5	32.4
✓	✓	✓	✓	**74.5**	**66.9**	**78.4**	**33.7**

Black bold numbers refer to top performance.

where TP denotes correctly recognized objects in the image, FP signifies incorrectly recognized objects, and FN indicates objects that were correctly recognized but assigned to an incorrect category.

AP refers to the area under the *P–R* curve, while the mAP represents the average value of AP for each category. Specifically, mAP@0.5 denotes the average value of AP when the intersection over union (IoU) threshold is set to 0.5. On the other hand, mAP@0.5:0.95 indicates the average mAP across different IoU thresholds (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95). The metrics AP and mAP are computed using the following equations:

$$\mathrm{AP} = \sum_{k=1}^{N} P(K)\Delta R(K) \tag{15}$$

$$m\mathrm{AP} = \frac{1}{C}\sum_{k=1}^{N} P(K)\Delta R(K) \tag{16}$$

$$\Delta R(K) = R(K) - R(K-1) \tag{17}$$

where C denotes the number of object categories, K represents the current IoU threshold, and $P(K)$ and $R(K)$ denote precision and recall of the current IoU threshold, respectively.

FLOPs serve as a metric for quantifying the computational complexity of a model and are frequently employed as an indirect indicator of the speed of a neural network model. On the other hand, Params represents the number of model parameters. In addition, FPS provides a measure of the efficiency of a model.

D. Dense Object Detection

1) Ablation Studies: To assess the effectiveness of the DS-YOLOv5s in enhancing dense object detection performance in UAV images, a series of ablation experiments are conducted on the VisDrone-2019 dataset and the results are shown in Table II. The findings reveal that the introduction of SIE on top of YOLOv5s resulted in a 0.3% increase in mAP@0.5 and a 0.8% increase in mAP@0.5:0.95, effectively improving both feature extraction capability and detection accuracy. In addition, the incorporation of Coord_C3 based on YOLOv5s yielded notable improvements across mAP@0.5 and mAP@0.5:0.95. Specifically, there was a 2% increase in mAP@0.5:0.95, significantly enhancing the accuracy of dense object detection. Similarly, the inclusion of EfficiCLoss in YOLOv5s led to enhancements in mAP@0.5 and mAP@0.5:0.95, indicating improved detection accuracy. DS-YOLOv5s signifies the adoption of a comprehensive approach that incorporates all three proposed improvement methods. The results unequivocally demonstrate the superior detection accuracy of the proposed method for dense objects in UAV images.

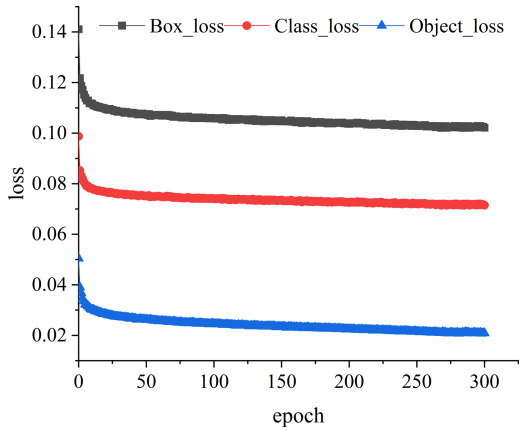

Fig. 10. Training loss on VisDrone-2019 dataset.

2) Comparison With State-of-the-Art Methods: The training process was conducted for a total of 300 epochs, and the corresponding loss curve is illustrated in Fig. 10. In this figure, Box_loss represents the discrepancy between the predicted and actual bounding boxes. Class_loss indicates the classification loss, which determines the model's ability to accurately recognize objects in the image and assign them to the correct categories. Object_loss represents the confidence loss, which supervises the presence of objects within the grid and calculates the network's confidence level. As shown in Fig. 10, it can be observed that after 300 epochs, the loss values of the DS-YOLOv5s network ceased to decrease, indicating that the network had converged and stabilized.

DS-YOLOv5s was compared with other object detection methods on the VidDrone-2019 dataset. As shown in Table III, on mAP@0.5, DS-YOLOv5 improves upon Faster RCNN by 36.3%, Mask RCNN by 16.8%, Cascade RCNN by 4.1%, CornerNet by 22.5%, CenterNet by 1.75%, SSD by 30.4%, YOLOv2 by 28%, YOLOv3 by 4%, YOLOV4 by 5.7%, YOLOv5s by 3.1%, EfficientDet by 2.6%, SGMFNet by 2.8%, YOLOv7-sea by 2%, and RTD-Net by 3.5%. On Params, YOLOV5s has the

TABLE III
COMPARISON RESULTS OF DIFFERENT METHODS ON VISDRONE-2019

Method	AP(%)	mAP@0.5 (%)	Params (M)	FLOPs (G)
Faster RCNN [54]	70.8(0.08)	42.3	41.16	299.2
Mask RCNN [55]	75.4(0.03)	62.1	60.7	310.57
Cascade RCNN [56]	79.1(0.09)	74.8	68.95	320.07
CornerNet [57]	69.4(0.11)	56.4	200.96	**1104.06**
CenterNet [58]	76.8(0.02)	77.2	70.75	137.21
SSD [59]	73.5(0.04)	48.5	24.39	175.2
YOLOv2 [60]	75.4(0.11)	50.9	48.7	19.8
YOLOv3 [61]	68.6(0.06)	74.9	8.5	21.1
YOLOv4 [62]	69.8(0.07)	73.2	9.42	20.4
YOLOv5s [23]	71.2(0.13)	75.8	**6.7**	15.8
EfficientDet [63]	74.8(0.10)	76.3	52	1.9
SGMFNet [64]	73.9(0.11)	76.1	62.3	305.8
YOLOv7-sea [65]	74.2(0.08)	76.9	55.7	283.2
RTD-Net [22]	73.6(0.05)	75.4	23.8	165.7
DS-YOLOv5s	**82.3(0.01)**	**78.9**	6.8	16.3

Black bold numbers refer to top performance.
(·) represents standard deviation.

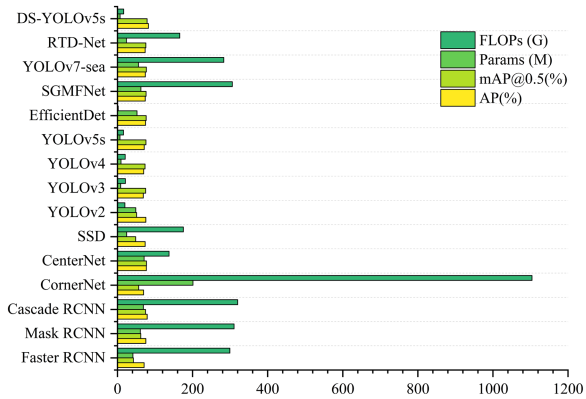

Fig. 11. Qualitative results on VisDrone-2019 dataset.

TABLE IV
COMPARISON OF OBJECT DETECTION PERFORMANCE OF VARIOUS CATEGORIES ON VISDRONE-2019(MAP@0.5(%))

Category	YOLOv5s [23]	DS-YOLOv5s	Variation
pedestrian	50.8	55.4	+4.6
people	44.7	51.3	+6.6
bicycle	35.4	40.8	+5.4
car	69.8	74.9	+5.1
van	44.2	49.8	+5.6
truck	42.9	46.4	+3.5
tricycle	22.9	28.7	+5.8
awning-tricycle	23.9	27.9	+4
bus	48.9	53.4	+4.5
motor	40.5	45.9	+5.4

The symbol "+" denotes the improvements of DS-YOLOv5s over YOLOv5s.

smallest number of parameters and our method is second only to YOLOv5s. On standard deviation, DS-YOLOv5s has the smallest standard deviation value of AP, which demonstrates the stability of the method. Due to the limitations of UAVs hardware, small platforms are more sensitive to the number of parameters and model volume. DS-YOLOv5s effectively reduces the number of parameters and volume of the network by using the idea of CoordConv and replacing the traditional convolution module of YOLOv5s with the CoordConv module. It can be seen from Table III and Fig. 11 that DS-YOLOv5s greatly reduces the hardware requirements of the network structure while guaranteeing accuracy, which is conducive to its use in small devices, such as UAVs.

As shown in Table IV, compared to YOLOv5s, mAP@0.5 of DS-YOLOv5s increases by 4.6%, 6.6%, 5.4%, 5.1%, 5.6%, 3.5%, 5.8%, 4%, 4.5%, and 5.4%, respectively, for the pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor categories. The experimental results show that DS-YOLOv5s can significantly improve the detection accuracy of the network for dense objects.

In addition, to further demonstrate the effectiveness of the proposed method, we conducted a comparative test with state-of-the-art methods on Stanford Drone dataset. As shown in Table V, on AP, DS-YOLOv5 improves upon Faster RCNN by 11.9%, Mask RCNN by 7.3%, Cascade RCNN by 3.6%, CornerNet by 13%, CenterNet by 5.9%, SSD by 9.3%, YOLOv2 by 6.5%, YOLOv3 by 14.1%, YOLOV4 by 12.4%, YOLOv5s by 11.4%, EfficientDet by 7.9%, SGMFNet by 8.6%, YOLOv7-sea by 8.1%, and RTD-Net by 9%. It can be seen from Table V and Fig. 12 that DS-YOLOv5s greatly reduces the hardware

TABLE V
COMPARISON RESULTS OF DIFFERENT METHODS ON STANFORD DRONE

Method	AP(%)	mAP@0.5 (%)
Faster RCNN [54]	71.3(0.08)	42.9
Mask RCNN [55]	75.9(0.03)	62.8
Cascade RCNN [56]	79.6(0.11)	75.3
CornerNet [57]	70.2(0.13)	56.2
CenterNet [58]	77.3(0.04)	77.5
SSD [59]	73.9(0.04)	49.1
YOLOv2 [60]	76.7(0.10)	50.6
YOLOv3 [61]	69.1(0.05)	72.3
YOLOv4 [62]	70.8(0.04)	75.6
YOLOv5s [23]	71.8(0.10)	77.1
EfficientDet [63]	75.3(0.09)	76.8
SGMFNet [64]	74.6(0.11)	77.3
YOLOv7-sea [65]	75.1(0.07)	75.8
RTD-Net [22]	74.2(0.05)	74.9
DS-YOLOv5s	**83.2(0.02)**	**80.1**

Black bold numbers refer to top performance.
(·) represents standard deviation.

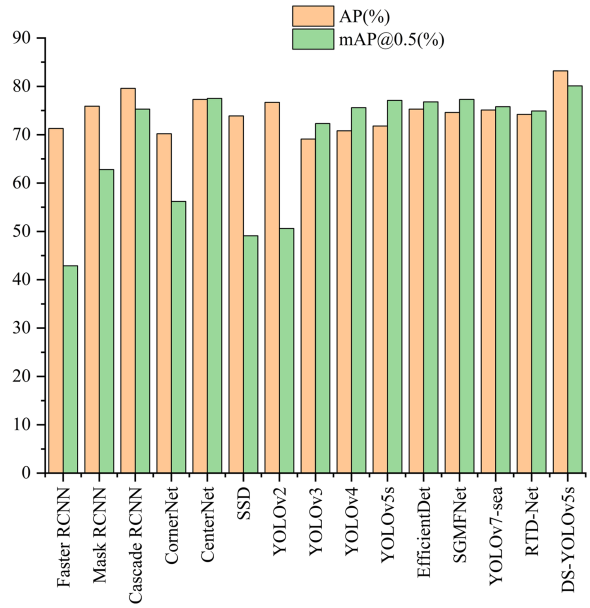

Fig. 12. Qualitative results on Stanford Drone dataset for dense objects.

requirements of the network structure while guaranteeing accuracy, which is conducive to the its use in small devices, such as UAVs.

3) Analysis of Visualization Results: To more intuitively demonstrate the dense object detection capability of DS-YOLOv5s in UAV images, we selected some experimental results, as shown in Fig. 13, which demonstrate that the proposed method can accurately determine the position of an object in dense objects. This is a challenging scenario, as the algorithm can easily misidentify these as a single object, or miss some of them. DS-YOLOv5s can effectively detect each object and

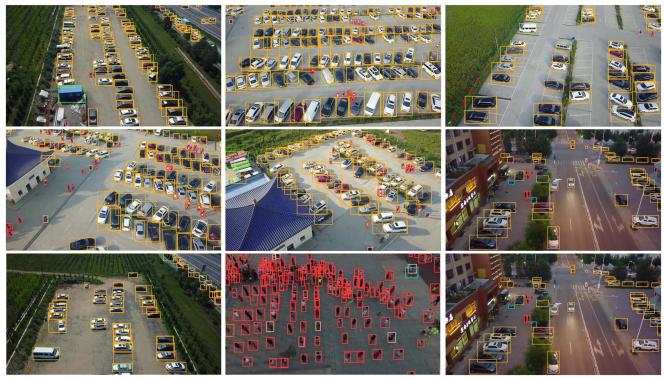

Fig. 13. Detection results of dense objects on VisDrone-2019.

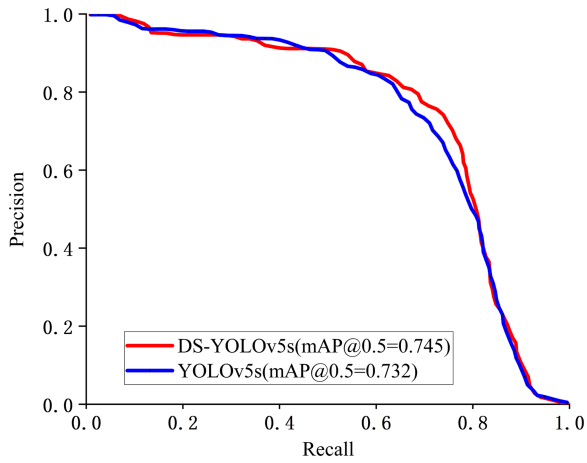

Fig. 14. P-R curve of YOLOv5s and DS-YOLOv5s on LEVIR-ship.

accurately recognize its location and category, showing strong robustness and accuracy.

E. Small Object Detection

1) Ablation Studies: Ablation experiments were conducted on the LEVIR-ship dataset to validate the performance of the proposed improvement measures for small object detection. As shown in Table VI, compared to the YOLOv5s, DS-YOLOv5s exhibits improvements of 1.8% in mAP@0.5 and 9% in mAP@0.5:0.95. These findings validate the effectiveness of the proposed module.

2) Comparison With State-of-the-Art Methods: Following 300 epochs of training, a test was conducted using sample data from the test dataset. The area under the P–R curve serves as a measure of average accuracy, with a larger area indicating higher accuracy. As shown in Fig. 14, the P–R curve of DS-YOLOv5s exhibits a larger area compared to that of YOLOv5s, indicating superior performance.

Experimental comparisons were conducted between DS-YOLOv5s and other object detection methods using the LEVIR-ship and Stanford Drone datasets. As shown in Table VII and Fig. 15, DS-YOLOv5s has the highest mAP@0.5 and mAP@0.5:0.95, which indicates the superior performance of

TABLE VI
ABLATION EXPERIMENTS ON LEVIR-SHIP

YOLOv5s [23]	SIE	Coord_C3	EfficiCLoss	mAP@0.5(%)	mAP@0.5:0.95(%)
✓				73.6	59.1
✓	✓			73.8	61.2
✓		✓		74.3	62.3
✓			✓	74.9	63.4
✓	✓	✓		75.1	64.9
✓	✓	✓	✓	**75.4**	**68.1**

Black bold numbers refer to top performance.

TABLE VII
COMPARISON OF ACCURACY OF DIFFERENT METHODS ON LEVIR-SHIP

Methods	mAP@0.5(%)	mAP@0.5:0.95(%)
Faster RCNN [54]	44.3	39.7
Mask RCNN [55]	52.9	48.6
Cascade RCNN [56]	61.5	53.5
CornetNet [57]	50.8	52.8
CenterNet [58]	68.4	63.9
SSD [59]	50.2	30.3
YOLOv2 [60]	53.8	39.4
YOLOv3 [61]	75.1	63.5
YOLOv4 [62]	73.2	64.1
YOLOv5s [23]	73.6	64.5
EfficientDet [63]	73.9	66.7
SGMFNet [64]	74.1	67.2
YOLOv7-sea [65]	74.4	67.9
RTD-Net [22]	73.8	66.5
DS-YOLOv5s	**75.4**	**68.1**

Black bold numbers refer to top performance.

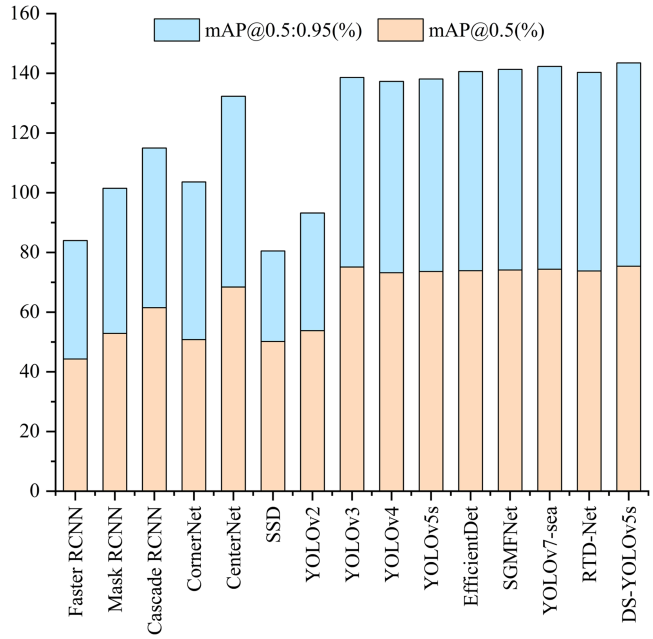

Fig. 15. Comparison of mAP@0.5 and mAP@0.5:0.95 on LEVIR-ship.

the method. In terms of mAP@0.5, DS-YOLOv5 outperforms Faster RCNN by 31.1%, Mask RCNN by 22.5%, Cascade RCNN by 13.9%, CornerNet by 24.6%, CenterNet by 7%, SSD by 25.2%, YOLOv2 by 21.6%, YOLOv3 by 0.3%, YOLOv4 by 2.2%, YOLOv5s by 1.8%, and EfficientDet by 1.5%. In addition, for mAP@0.5:0.95, DS-YOLOv5 surpasses Faster RCNN by 28.4%, Mask RCNN by 19.5%, Cascade RCNN by 14.6%, CornerNet by 15.3%, CenterNet by 4.2%, SSD by 37.8%, YOLOv2 by 28.7%, YOLOv3 by 4.6%, YOLOv4 by 4%, YOLOv5s by 3.6%, and EfficientDet by 1.4%. These findings highlight the efficiency and effectiveness of our proposed method.

As shown in the Table VIII and Fig. 16, the accuracy on Stanford Drone is higher than on the LEVIR-ship dataset compared to Table VII. In terms of mAP@0.5:0.95, DS-YOLOv5 outperforms Faster RCNN by 35.1%, Mask RCNN by 26.2%, Cascade RCNN by 19.7%, CornerNet by 26.6%, CenterNet by 9.6%, SSD by 16.8%, YOLOv2 by 24.1%, YOLOv3 by 2.5%, YOLOv4 by 5.3%, YOLOv5s by 4.8%, EfficientDet by 3%, SGMFNet by 1.5%, YOLOv7-sea by 0.8%, and RTD-Net by 3.3%. At the same time, YOLOv7-sea has the second best result. Experimental results on LEVIR-ship and Stanford Drone datasets demonstrate the superiority as well as accuracy of DS-YOLOv5s for small object detection.

3) Analysis of Visualization Results: To assess the generalization capability of DS-YOLOv5s in detecting small objects, we conducted comparative experiments using the LEVIR-ship dataset. As shown in Fig. 17, DS-YOLOv5s is able to accurate detection the smallest objects almost without missed or false detection.

F. Computational Complexity

Table IX presents the FPS achieved by each object detection method on the VisDrone-2019, LEVIR-ship, and Stanford Drone datasets using the same experimental platform. The results reveal that Faster-RCNN exhibits the lowest FPS due to its two-stage network architecture. This approach involves extracting the object region first, followed by CNN classification and identification, which yields higher detection accuracy but can hardly meet the real-time requirement. On the other hand, YOLOv5s achieves the highest FPS but with lower accuracy. In terms of

TABLE VIII
COMPARISON OF ACCURACY OF DIFFERENT METHODS ON STANFORD DRONE FOR SMALL OBJECTS

Methods	mAP@0.5(%)	mAP@0.5:0.95(%)
Faster RCNN [54]	45.7	41.8
Mask RCNN [55]	53.6	50.7
Cascade RCNN [56]	62.1	57.2
CornetNet [57]	51.9	50.3
CenterNet [58]	68.7	67.3
SSD [59]	50.8	50.1
YOLOv2 [60]	54.3	52.8
YOLOv3 [61]	75.6	74.4
YOLOv4 [62]	73.8	71.6
YOLOv5s [23]	74.2	72.1
EfficientDet [63]	75.1	73.9
SGMFNet [64]	76.2	75.4
YOLOv7-sea [65]	76.8	76.1
RTD-Net [22]	75.7	73.6
DS-YOLOv5s	**78.4**	**76.9**

Black bold numbers refer to top performance.

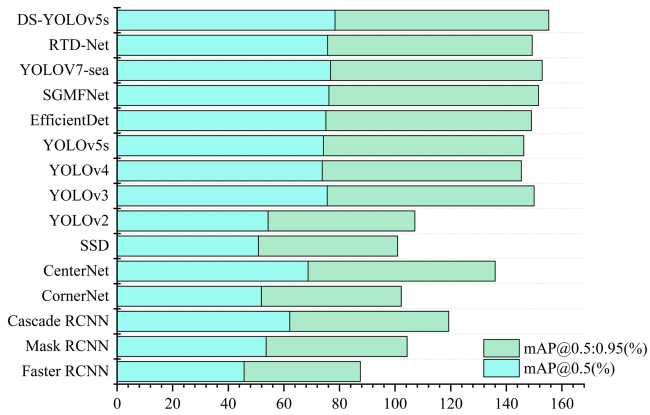

Fig. 16. Comparison of mAP@0.5 and mAP@0.5:0.95 on Stanford Drone for small objects.

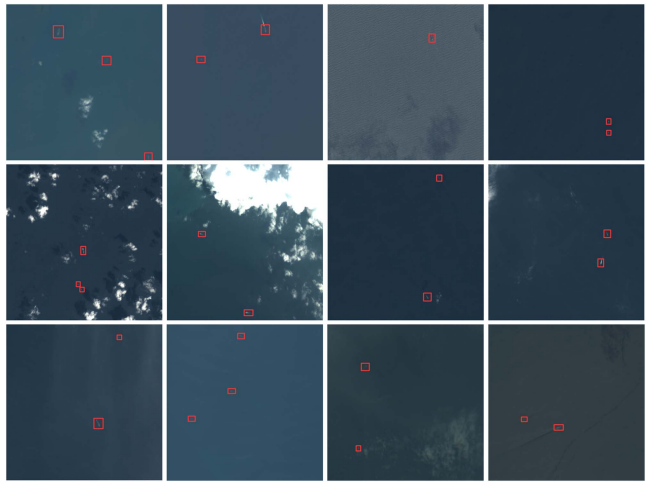

Fig. 17. Detection results of small objects on LEVIR-ship (red rectangular boxes indicate detection of small objects).

TABLE IX
COMPARISON OF FPS ON VISDRONE-2019, LEVIR-SHIP AND STANFORD DRONE DATASET

Methods	VisDrone-2019	LEVIR-ship	Stanford Drone
Faster RCNN [54]	4.9	5.5	6.2
Mask RCNN [55]	5.8	6.1	7.9
Cascade RCNN [56]	10.9	11.5	11.7
CornetNet [57]	13.1	14.7	15.3
CenterNet [58]	13.5	13.8	16.1
SSD [59]	16.7	17.2	17.8
YOLOv2 [60]	47.9	48.4	50.1
YOLOv3 [61]	59.1	59.7	60.6
YOLOv4 [62]	55.3	56.4	58.2
YOLOv5s [23]	**60.2**	**63.7**	**65.4**
EfficientDet [63]	34.8	35.1	35.7
SGMFNet [64]	48.3	49.1	49.8
YOLOv7-sea [65]	46.2	47.9	48.3
RTD-Net [22]	22.4	23.8	23.9
DS-YOLOv5s	50.2	51.5	52.7

Black bold numbers refer to top performance.

FPS, DS-YOLOv5s outperforms YOLOv3 and YOLOv4 on the VisDrone-2019, LEVIR-ship and Stanford Drone datasets, with FPS values of 50.2, 51.5, and 52.7, respectively. The results demonstrate that DS-YOLOv5s has higher detection accuracy and faster detection speed, which shows that DS-YOLOv5s is a promising object detection method for dense and small objects in UAV applications.

V. DISCUSSION

With the popularity and development of UAV technology, UAVs are widely used in military and civilian applications. However, due to the high flight altitude and large field of view of UAVs, the images captured from the UAV contain both dense and small objects, which reduces the accuracy of object detection by UAVs. Therefore, it is crucial to develop a method that can simultaneously detect dense and small objects in UAV images.

Existing UAV object detection methods are mainly object-specific and have low accuracy for detecting dense and small objects in images. In addition, existing methods suffer from large number of parameters and low computational efficiency, which make them difficult to be deployed on UAV computing platforms and to perform real-time object detection. Our proposed method is able to achieve real-time object detection while guaranteeing high-accuracy detection of dense and small objects. First, a kernel K-mean clustering algorithm is used to optimize the anchors of the dataset. Then, SIE module is introduced in the backbone and neck of the network to enhance the location information

of dense objects in the network by extracting spatial feature information. In the backbone network, based on CoordConv, the Coord_C3 module is proposed to replace the C3 module of YOLOv5s, which improves the acceptance domain of the network and reduces the number of parameters of the network. In addition, in the head of the network, a detection head for small objects is added to improve the detection accuracy of small objects. Finally, the introduction of jump connection fusion of shallow and deep features improves the feature sensing ability of the network, which further improves the detection accuracy of dense small objects.

It should be noted that, the proposed model is still large to be deployed on UAVs and the detection performance also needs to improve considering the complexity in real applications. More specifically:

1) proposing a UAV image object detection method under hazy weather to improve robustness in complex environments;
2) proposing a dynamic object detection method for UAV images to achieve real-time object tracking;
3) improving the computational resources of UAV computing platforms.

VI. Conclusion

This article focuses on improving the detection accuracy of dense and small objects in UAV images by incorporating feature fusion and spatial information. To this end, we have proposed DS-YOLOv5s. First, we introduce the kernel K-means algorithm at the Input of the network, enabling rapid determination of the optimal anchor size. To effectively extract spatial information from the feature map, an SIE module is proposed. Furthermore, Coord_C3 module is introduced to extend the range of feature awareness and to reduce the model size. In addition, skip connections are employed to fuse shallow strong semantic information with deep weak semantic information, thereby enhancing the network's receptive field. Finally, we incorporate a small object detection head into the network architecture to improve the detection of small objects. Experimental results demonstrate that DS-YOLOv5s surpasses existing state-of-the-art methods in terms of both detection accuracy and FPS.

As future work, we plan to evaluate our method with more challenging datasets and design a lightweight network to reduce the size of the network.

References

[1] J. Han et al., "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogramm. Remote Sens.*, vol. 89, pp. 37–48, 2014.

[2] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.

[3] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Apr. 2017.

[4] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2844–2853.

[5] X. Han, Y. Zhong, and L. Zhang, "An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery," *Remote Sens.*, vol. 9, no. 7, 2017, Art. no. 666.

[6] W. Sun, L. Dai, X. Zhang, P. Chang, and X. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Appl. Intell.*, vol. 52, no. 8, pp. 8448–8463, 2021.

[7] 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.

[8] J. Wang, F. Shao, X. He, and G. Lu, "A novel method of small object detection in UAV remote sensing images based on feature alignment of candidate regions," *Drones*, vol. 6, no. 10, 2022, Art. no. 292.

[9] J. Zhang, G. Wan, M. Jiang, G. Lu, X. Tao, and Z. Huang, "Small object detection in UAV image based on improved YOLOv5," *Syst. Sci. Control. Eng.*, vol. 11, no. 1, 2023, Art. no. 2247082.

[10] S. Zeng, W. Yang, Y. Jiao, L. Geng, and X. Chen, "SCA-YOLO: A new small object detection model for UAV images," *Vis Comput.*, vol. 40, pp. 1787–1803, 2024.

[11] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "UAV-YOLO: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, 2020, Art. no. 2238.

[12] H. Zhou, A. Ma, Y. Niu, and Z. Ma, "Small-object detection for UAV-based images using a distance metric method," *Drones*, vol. 6, no. 10, 2022, Art. no. 308.

[13] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501212.

[14] Y. Liu, F. Yang, and P. Hu, "Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks," *IEEE Access*, vol. 8, pp. 145740–145750, 2020.

[15] W. Xu, C. Zhang, Q. Wang, and P. Dai, "FEA-Swin: Foreground enhancement attention Swin transformer network for accurate UAV-based dense object detection," *Sensors*, vol. 22, no. 18, 2022, Art. no. 6993.

[16] J. Feng, Y. Liang, X. Zhang, J. Zhang, and L. Jiao, "SDANet: Semantic-embedded density adaptive network for moving vehicle detection in satellite videos," *IEEE Trans. Image Process.*, vol. 32, pp. 1788–1801, 2023.

[17] T. Ye, W. Qin, Y. Li, S. Wang, J. Zhang, and Z. Zhao, "Dense and small object detection in UAV-vision based on a global-local feature enhanced network," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2515513.

[18] P. Zhang, Y. Zhong, and X. Li, "SlimYOLOv3: Narrower, faster and better for real-time UAV applications," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 37–45.

[19] K. Gromada, B. Siemiątkowska, W. Stecz, K. Płochocki, and K. Woźniak, "Real-time object detection and classification by UAV equipped with SAR," *Sensors*, vol. 22, no. 5, 2022, Art. no. 2068.

[20] C. Kyrkou, G. Plastiras, T. Theocharides, S. I. Venieris, and C.-S. Bouganis, "DroNet: Efficient convolutional neural network detector for real-time UAV applications," in *Proc. Des. Autom. Test Europe Conf. Exhib.*, 2018, pp. 967–972.

[21] Z. Zhang, Y. Liu, T. Liu, Z. Lin, and S. Wang, "DAGN: A real-time UAV remote sensing image vehicle detection framework," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 11, pp. 1884–1888, Nov. 2020.

[22] T. Ye, W. Qin, Z. Zhao, X. Gao, X. Deng, and Y. Ouyang, "Real-time object detection network in UAV-vision based on CNN and transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 2505713.

[23] G. Jocher et al., "ultralytics/yolov5: V3. 0," Zenodo, 2020.

[24] 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.*, 2016, pp. 779–788.

[25] F. Dadboud, V. Patel, V. Mehta, M. Bolic, and I. Mantegh, "Single-stage UAV detection and classification with YOLOV5: Mosaic data augmentation and PANet," in *Proc. 17th IEEE Int. Conf. Adv. Video Signal-Based Surveill.*, 2021, pp. 1–8.

[26] M. Gao, Y. Du, Y. Yang, and J. Zhang, "Adaptive anchor box mechanism to improve the accuracy in the object detection system," *Multimedia Tools Appl.*, vol. 78, pp. 27383–27402, 2019.

[27] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 390–391.

[28] K. He, X. Zhang, S. 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, Sep. 2015.

[29] Y. Zhang, J. H. Han, Y. W. Kwon, and Y. S. Moon, "A new architecture of feature pyramid network for object detection," in *Proc. 6th IEEE Int. Conf. Comput. Commun.*, 2020, pp. 1224–1228.
[30] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.
[31] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IOU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12993–13000.
[32] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit.*, 2006, pp. 850–855.
[33] K. Jin, Y. Chen, B. Xu, J. Yin, X. Wang, and J. Yang, "A patch-to-pixel convolutional neural network for small ship detection with PolSAR images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 58, no. 9, pp. 6623–6638, Sep. 2020.
[34] J. Han, S. Moradi, I. Faramarzi, C. Liu, H. Zhang, and Q. Zhao, "A local contrast method for infrared small-target detection utilizing a tri-layer window," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 10, pp. 1822–1826, Oct. 2020.
[35] J. Wang, Y. Lin, J. Guo, and L. Zhuang, "SSS-YOLO: Towards more accurate detection for small ships in SAR image," *Remote. Sens. Lett.*, vol. 12, no. 2, pp. 93–102, 2021.
[36] L. Yu, H. Wu, Z. Zhong, L. Zheng, Q. Deng, and H. Hu, "TWC-Net: A SAR ship detection using two-way convolution and multiscale feature mapping," *Remote Sens.*, vol. 13, no. 13, 2021, Art. no. 2558.
[37] W. Li, J. Wang, Y. Gao, M. Zhang, R. Tao, and B. Zhang, "Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[38] J. Zhao, W. Guo, Z. Zhang, and W. Yu, "A coupled convolutional neural network for small and densely clustered ship detection in SAR images," *Sci. China Inf. Sci.*, vol. 62, pp. 1–16, 2019.
[39] W. Dai, Y. Mao, R. Yuan, Y. Liu, X. Pu, and C. Li, "A novel detector based on convolution neural networks for multiscale sar ship detection in complex background," *Sensors*, vol. 20, no. 9, 2020, Art. no. 2547.
[40] J. Wang, W. Li, M. Zhang, and J. Chanussot, "Large kernel sparse convnet weighted by multi-frequency attention for remote sensing scene understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5626112.
[41] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2020.
[42] P. Xu et al., "On-board real-time ship detection in HISEA-1 SAR images based on CFAR and lightweight deep learning," *Remote Sens.*, vol. 13, no. 10, 2021, Art. no. 1995.
[43] Y. Liang, J. Feng, X. Zhang, J. Zhang, and L. Jiao, "Midnet: An anchor-and-angle-free detector for oriented ship detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612113.
[44] R. Liu et al., "An intriguing failing of convolutional neural networks and the CoordConv solution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–12.
[45] K. Krishna and M. N. Murty, "Genetic k-means algorithm," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 29, no. 3, pp. 433–439, Mar. 1999.
[46] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," *Proc. VLDB Endowment (PVLDB)*, vol. 5, no. 7, pp. 622–633, 2012.
[47] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: Spectral clustering and normalized cuts," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2004, pp. 551–556.
[48] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, 2022.
[49] Z. Gevorgyan, "SIoU loss: More powerful learning for bounding box regression," 2022, *arXiv:2205.12740*.
[50] P. Zhu et al., "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, Nov. 2022.
[51] J. Chen, K. Chen, H. Chen, Z. Zou, and Z. Shi, "A degraded reconstruction enhancement-based method for tiny ship detection in remote sensing images with a new large-scale dataset," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 5625014.
[52] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.
[53] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Stat.*, vol. 22, no. 3, pp. 400–407, 1951.
[54] 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.
[55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[56] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
[57] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.
[58] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6568–6577.
[59] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[60] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6517–6525.
[61] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.
[62] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOV4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[63] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10778–10787.
[64] Y. Zhang, C. Wu, T. Zhang, Y. Liu, and Y. Zheng, "Self-attention guidance and multiscale feature fusion-based UAV image object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6004305.
[65] H. Zhao, H. Zhang, and Y. Zhao, "Yolov7-Sea: Object detection of maritime UAV images based on improved YOLOv7," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops*, 2023, pp. 233–238.

Chenyang Li (Student Member, IEEE) received the B.E. degree in detection guidance and control techniques from the North University of China, Taiyuan, China, in 2019. He is currently working toward the Ph.D. degree in instruments science and technology with the School of Aerospace Science and Technology, Xidian University, Xi'an, China.

His research interests include computer vision, image processing, and UAV object detection.

Suiping Zhou received the B.E., M.A., and Ph.D. degrees in electrical engineering from Beihang University, Beijing, China, in 1989, 1992, and 1995, respectively.

He was a Full Professor with the Department of Computer Science, Middlesex University, London, U.K., and an Assistant Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. He is currently a Full Professor with the School of Aerospace Science and Technology, Xidian University, Xi'an, China. His research interests include large-scale distributed interactive applications, parallel/distributed systems, and Big Data in space science and technology.

Hang Yu (Member, IEEE) received the B.E. degree in electronic and information engineering and the Ph. D. degree in circuits and systems from Xidian University, Xi'an, China, 2004 and 2015, respectively.

He is currently an Associate Professor with the School of Aerospace Science and Technology, Xidian University. His research interests include image understanding and interpretation, computer vision, machine learning, information fusion, evolutionary computation and optimization, visual navigation, and SAR image processing.

Tianxiang Guo received the B.E. degree in automation and the M.S. degree in pattern recognition and intelligent systems from Hebei University, Baoding, China, 2018 and 2021, respectively.

He is currently working toward the Ph.D. degree in instruments science and technology with the School of Aerospace Science and Technology, Xidian University, Xi'an, China. His research interest includes medical image analysis, especially in noninvasive stimulation with machine learning methods

Yuru Guo received the B.E. degree in detection guidance and control techniques from the North University of China, Taiyuan, China, in 2019. She is currently working toward the M.S. degree in electronic information engineering with the School of Aerospace Science and Technology, Xidian University, Xi'an, China.

Her research interest includes computer vision, image processing, and remote sensing.

Jichen Gao received the B.E. degree in electronic information engineering from Yanshan University, Qinhuangdao, China, in 2022. He is currently working toward the M.S. degree in electronic information with the School of Aerospace Science and Technology, Xidian University, Xi'an, China.

His research interests include object detection, image processing, artificial intelligence, and UAV.