Journal of Systems Engineering and Electronics
Vol. 32, No. 4, August 2021, pp.841 – 853

Low-altitude small-sized object detection using lightweight feature-enhanced convolutional neural network

YE Tao[1,*], ZHAO Zongyang[1], ZHANG Jun[1], CHAI Xinghua[2], and ZHOU Fuqiang[3]

1. School of Mechanical and Electrical Information Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China; 2. The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China; 3. School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing 100083, China

Abstract: Unauthorized operations referred to as "black flights" of unmanned aerial vehicles (UAVs) pose a significant danger to public safety, and existing low-attitude object detection algorithms encounter difficulties in balancing detection precision and speed. Additionally, their accuracy is insufficient, particularly for small objects in complex environments. To solve these problems, we propose a lightweight feature-enhanced convolutional neural network able to perform detection with high precision detection for low-attitude flying objects in real time to provide guidance information to suppress black-flying UAVs. The proposed network consists of three modules. A lightweight and stable feature extraction module is used to reduce the computational load and stably extract more low-level feature, an enhanced feature processing module significantly improves the feature extraction ability of the model, and an accurate detection module integrates low-level and advanced features to improve the multiscale detection accuracy in complex environments, particularly for small objects. The proposed method achieves a detection speed of 147 frames per second (FPS) and a mean average precision (mAP) of 90.97% for a dataset composed of flying objects, indicating its potential for low-altitude object detection. Furthermore, evaluation results based on microsoft common objects in context (MS COCO) indicate that the proposed method is also applicable to object detection in general.

Keywords: unmanned aerial vehicle (UAV), deep learning, lightweight network, object detection, low-attitude.

DOI: 10.23919/JSEE.2021.000073

1. Introduction

Unmanned aerial vehicles (UAVs) have been widely adopted in a variety of industrial, consumer, and military applications with significant effects on society as a whole, owing to their unique capabilities in terms of improving national defense [1] and convenience for civilian use. However, as yet UAV-related laws have not been perfected in existing legislation, the phenomenon of black flights occurs frequently, violating personal privacy and public safety, or even national security [2]. Given the increasing number of UAVs (particularly small UAVs), it is becoming increasingly difficult to monitor them effectively in complex or low-light environments. In addition, it remains challenging to guarantee accurate detection of UAVs under conditions of abundant low-altitude interfering objects such as birds and kites. Therefore, to protect public property, provide effective air traffic control information, and suppress the phenomena of black flights, it is essential to develop methods able to perform highly accurate real-time detection of UAVs under a wide variety of environmental conditions.

A variety of approaches to detecting low-flying objects have been developed in recent years [3−6]. However, existing object-detection algorithms cannot guarantee the detection of multiscale objects in dark and complex environments, especially small objects. Currently existing methods have significant difficulty in detecting objects of small volumes at low altitudes, in poor lighting without sufficient extracted features. Additionally, low-altitude object-detection algorithms should perform both accurately and rapidly, which is crucial for accurately identifying unauthorized unmanned vehicles and allowing appropriate authorities to respond quickly. Hence, low-altitude object-detection algorithms must detect the various types and positions of objects precisely in real time.

To address these problems, we propose an end-to-end lightweight detection network architecture based on a fusion of multiscale features for the detection of small objects flying at low attitudes, called LSL-Net, based on YOLOv4-tiny. The network improves object detection performance in complex environments, particularly for small objects; moreover, it requires less detection time.

Manuscript received February 19, 2021.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (52075027) and the Fundamental Research Funds for the Central Universities (2020XJJD03).

The proposed method comprises three simple but significant modules, including a lightweight and stable feature extraction module (LSM), an enhanced feature processing module (EFM), and an accurate detection module (ADM). The LSM reduces the size of images and enhances information related to low-level features, while the EFM achieves a more powerful feature extraction by inserting a proposed new attention mechanism and a spatial pyramid pooling-network (SPP-Net) [7]. The ADM specially identifies a scale for more accurate detection of small objects in particular after the feature information is extracted by the lightweight network block (LNB). The proposed model, composed of these simple but effective modules, balances detection speed and accuracy well, and our experimental results demonstrate its excellent performance on the task of detecting low-altitude objects.

The contributions of this study are summarized as follows.

(i) We present an end-to-end multiscale lightweight object-detection network called LSL-Net offering a good balance between detection accuracy and speed, comprised of three modules, including LSM, EFM, and ADM. The LSM extracts stable low-level information and reduces the overall computational load, while the EFM extracts more effective information to perform more accurate detection, and the ADM achieves a higher precision for low-altitude objects at different scales, particularly small objects.

(ii) Techniques such as a cross-stage partial network (CSPNet) and an attention mechanism are used to improve detection precision. For an input image with a size of 416 × 416 pixels, our model achieves a mean average precision (mAP) of 90.97% and a detection speed of 147 frames per second (FPS). Comparative experiments show that the improvement in mAP over the benchmark methods is 6.71% higher than YOLOv4-tiny alone. Moreover, detection accuracy for UAV targets, is improved by 1.79% compared with prior methods, which indicates more accurate detection of small objects in low-altitude environments.

(iii) The experimental results show that LSL-Net demonstrates excellent performance in terms of balancing detection accuracy and speed. With an input size of 416 × 416 pixels, the forward inference speed is more than three times faster than that of a single-shot multibox detector (SSD), according to the experimental results on a low-attitude dataset. The proposed method demonstrates the ability to perform accurate real-time detection in diverse weather conditions and suppress UAV black flights.

The remainder of this study is organized as follows. Section 2 summarizes related work. Section 3 presents the proposed LSL-Net in detail, and Section 4 presents experimental results comparing the proposed network and various widely used models. Finally, Section 5 presents our conclusions.

2. Related works

2.1 Traditional object-detection algorithms

In recent years, research on object detection has been extensively conducted, and may be divided into traditional detection algorithms and methods based on deep learning, which has been extensively investigated. Most traditional detection algorithms achieve feature extraction and object-category detection with a combination of AdaBoost [8], the histogram of oriented gradients (HOG) [9] algorithm, and support-vector machines (SVMs) [10]. Nagahashi et al. [11] proposed parameterized AdaBoost, which achieved a faster training convergence by modifying parameters. However, its detection speed was insufficient, and its multi-class object detection accuracy in complex scenes could not be assured. Wang et al. [12] used AdaBoost to detect UAVs after extracting brightness, orientation, and regional contrast features. Omid-Zohoor et al. [13] used the illumination invariance of the HOG algorithm to enhance the detection capabilities of their proposed method. These two methods improve in detection accuracy; however, their real-time performance is poor as a result of excessive computational complexity, and their ability to detect small objects has not been verified. Liu et al. [14] analyzed motion characteristics and local features of small moving objects to achieve better detection of small UAVs based on a random forest method. Li et al. [15] designed an SVM-based detector by extracting three different features to strengthen its ability to detect small objects. Nevertheless, the types of extracted features overwhelmingly relied on the experience of designers, and room for improvement in the model's performance remained owing to the size limit of their datasets. Bazi et al. [16] exploited the state-of-the-art SVM to ensure precision of recognition for a limited number of training images. However, the recognition ability of this model in rough environments still remains to be tested. Overall, the generalization and robustness of these traditional algorithms largely do not meet the industrial requirements. These methods require large datasets to achieve high accuracy, so the high computational complexity results in poor real-time performance. Therefore, such methods are limited in their ability to realize accurate real-time detection in low-altitude environments.

2.2 Object-detection algorithms based on deep learning

Self-adaptive feature extraction algorithms based on convolutional neural networks (CNNs) [17] have achieved

promising successes with the development of deep learning, overcoming the feature extraction limitations of over-reliance on designers' experience. Among these algorithms, single-stage and two-stage versions have been developed. Two-stage methods, such as continuously improving region-CNN (R-CNN) series [18–21], generate candidate regions first and then perform object detection. Computationally expensive calculations significantly reduce their real-time performance, despite their high detection accuracy. Single-stage algorithms such as SSD [22] and YOLO [23–26] predict object positions and categories directly by means of regression, and thus they have more concise structures, and have demonstrated real-time detection capabilities. However, their detection accuracy remains low compared to two-stage methods. For example, the detection accuracies of CenterNet [27] and a full convolutional one-stage object detector (FCOS) [28] are higher than those of SSD and YOLO. However, their detection speed remains insufficient to meet the requirements of real-time detection. Li et al. [29] proposed two recognition algorithms for UAVs. They first pursued high detection accuracy and sent identified UAV images (detected by SSD) to AlexNet [30] for fine tuning. Nonetheless, their real-time performance could not be guaranteed in variable low-altitude environments. Another approach uses the k-nearest neighbors algorithm for classification after obtaining the position of a UAV with an SSD. The method assures real-time performance at the cost of low accuracy, especially for small targets. To significantly improve detection accuracy for small objects, deconvolutional single shot detector (DSSD) in [31] added deconvolutional and prediction modules based on SSD. The refined feature pyramid network-fully convolutional one-shot object detector in [32] extracted more abundant features by optimizing a feature pyramid network (FPN) structure. However, the detection speed decreases significantly compared to prior methods. Fan et al. [33] added a fully connected layer and a deconvolution layer to the SSD and performed UAV detection on low-resolution images. FII-CenterNet in [34] improved detection accuracy by introducing information on the location and scale of image foregrounds. These models are designed to balance detection accuracy and speed, whereas the ability to detect small objects in complex environments is not considered. Ma et al. [35] improved UAV detection stability by optimizing the ResNet block of YOLOv3. Cui et al. [36] adopted k-means clustering [37] to refine the anchors of YOLOv3 for a higher accuracy on UAV detection tasks. Although the detection speed is improved compared with SSDs, the real-time performance is insufficient efficiently to detect low-altitude UAVs. Wei et al. [38] improved the detection speed of YOLOv3 by reducing scales and concatenating features. However, the small object recognition ability of the model is nonetheless insufficient.

In summary, traditional methods for low-attitude object detection cannot self-adaptively extract or guarantee accurate identification in complex environments. However, object detection algorithms using deep learning can extract features adaptively. The detection results of deep learning-based approaches are generally better than those of traditional methods; however, it remains difficult to balance detection speed and accuracy. Additionally, the accuracy of automated systems detecting small objects at low altitudes should be improved to meet the evolving of industry requirements. To solve these problems, we propose a low-altitude small object detection model using a lightweight feature-enhanced CNN LSL-Net, realizing an excellent compromise between detection accuracy and speed, and improving object detection accuracy in harsh environments, particularly for small objects. The network is based on a light backbone consisting of group convolution and CSPNet [39] to balance the detection accuracyandcalculatecomputationalload.Ourexperimentalresults show that the proposed model can efficiently detect the movements of small, flying objects at low attitudes. Thus, the proposed method can achieve high-precision, real-time detection of UAVs in complex low-altitude environments.

3. Proposed network framework

We develop a new end-to-end adaptive feature information extraction and lightweight detection network inspired by YOLOv4-tiny, enabling high-precision real-time detection of flying objects at low attitudes. The network consists of three modules, including LSM, EFM, and ADM. LSM and EFM constitute the backbone of the network. The LSM reduces the computational load and improves detection speed and EFM fully extracts features and improves detection accuracy through an attention module, while ADM allows high-precision multiscale detection and strengths the system's ability to detect small objects.

3.1 Lightweight and stable feature extraction module

The low-level features of an image are invariably degraded to the point that they must be ignored when a feature map is significantly reduced. To prevent low-level feature information loss during image downsampling, the LSM is designed to perform better feature extraction by a muti-branch method and reduce the size of the input image. Its structure is shown in Fig. 1.

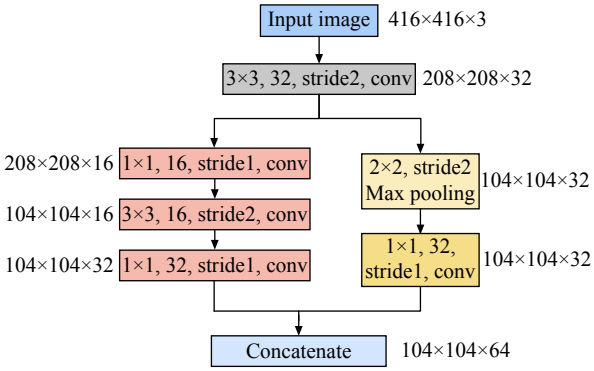


Fig. 1 LSM

In this module, the input image is reduced by a 3 × 3 convolution with two strides, and the image features are extracted at the same time. Multi-branch networks are then used to acquire different receptive fields. 1×1 convolutions are used to reduce the parameters and integrate information. Finally, we concatenate the information to prepare for the subsequent feature extraction step. To analyze the computations of the LSM and the traditional methods, floating point operations per second (FLOPs) are used to compute the computational complexity, which can be expressed as

$$\text{FLOPs} = \sum_{n=1}^{S} M_n^2 \cdot K_n^2 \cdot C_{n-1} \cdot C_n \tag{1}$$

where S represents the sum of convolutions, M_n^2 is the size of an output feature, K_n^2 is the kernel size, and C_{n-1} and C_n respectively represent the number of input channels and output channels. Hence, the FLOPs of the traditional convolution used in YOLOv4-tiny are given by

$$\text{FLOPs} = 104^2 \times 3^2 \times 32 \times 64 = 1.994 \times 10^8. \tag{2}$$

The FLOPs of the LSM is given by

$$\begin{aligned}\text{FLOPs} = {} & 208^2 \times 1^2 \times 32 \times 16 + 104^2 \times 3^2 \times 16 \times 16 + \\ & 104^2 \times 1^2 \times 16 \times 32 + 104^2 \times 32 \times 2^2 + 104^2 \times 1^2 \times 32 \times 32 + \\ & 208^2 \times 3^2 \times 32 \times 3 = 1.025 \times 10^8.\end{aligned} \tag{3}$$

Compared with the traditional low-level feature extraction method, the computational complexity of LSM is nearly reduced by two times which indicates that the LSM improves the feature-representation capability of the system and achieves stable downsampling without increasing the computational load. Taking multiple factors into consideration, such as the small objects in the dataset collected at low-attitude, as well as the accuracy and speed of detection, we obtain fixed-size images (such as 416 × 416) through data processing. By sending input data to the LSM before the EFM, we can enhance the feature extraction ability of the network without introducing additional calculations.

3.2 EFM

The feature extraction network model structure is a crucial factor in achieving good detection results. Many network models have been proposed, such as ResNet [40] and ResNext [41], which can achieve high accuracy or real-time detection in different scenarios.

However, in the context of the present work, the physical environment and the legal framework involved are complex and present considerable challenges to efficient detection. To achieve the subsequent high-precision object detection, the requirement of model feature-extraction ability is more stringent. According to the characteristics of the scenario under consideration, an EFM inspired by CSPNet is proposed, as shown in Fig. 2. The EFM is composed mainly of three CSP blocks and an attention module. Each CSP block achieves adequate feature extraction and reduces the size of the feature map by max-pooling. Furthermore, the attention module reorganizes the feature information to prepare for the powerful feature extraction of CSP-SPP.

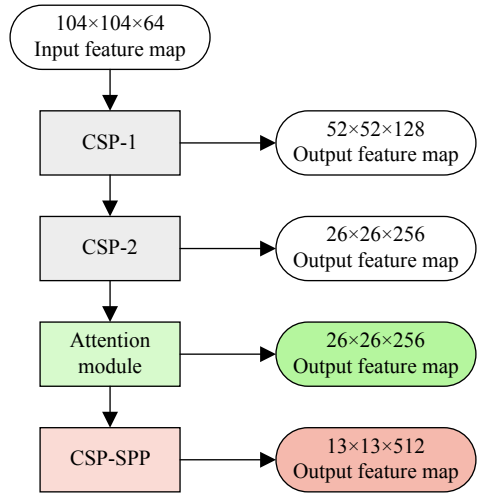


Fig. 2 EFM

The structures of the CSP-1 and CSP-2 layers are shown in Fig. 3. After a 3 × 3 convolution, the output channels are divided into two groups. To achieve a good compromise between detection accuracy and speed, the group of unprocessed feature maps is concatenated with the other group processed by various convolutions directly.

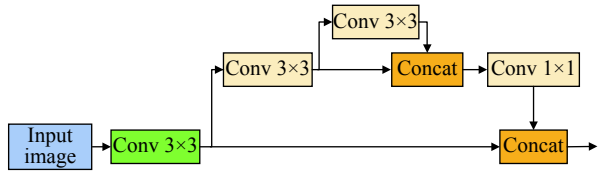


Fig. 3 Frame of CSP-1 and CSP-2

Attention mechanisms are designed to increase the weight of significant information and reduce that of unimportant information to improve the detection ability of computer vision models. Because various channels correspond to different responses, the channel attention mechanism shown in Fig. 4 is added before the final CSP-Net, encoding semantic information between channels to improve the detection accuracy.

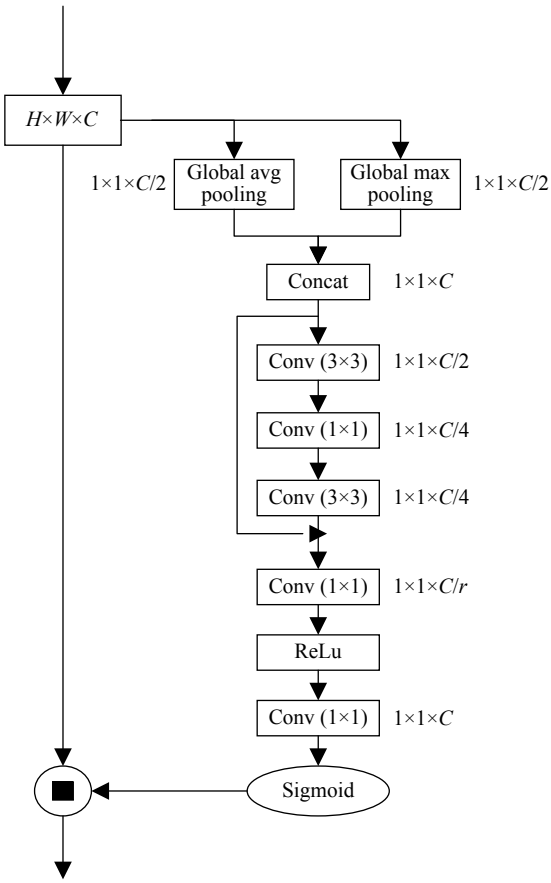


Fig. 4 Attention module

We first generate two different spatial context descriptors using both average- and max-pooling. Then, we concatenate the feature maps to integrate different information extracted by these pooling operations. A residual block is used to enhance the performance of this module, in which a 1×1 convolution is used to reduce the input feature dimensions, and a 3×3 convolution is adopted to enhance the feature expression ability of different channels. H and W represent the height and width of the feature map, respectively, while C and r respectively signify the channel and the ratio. To maximize the performance of the attention module performance, we set C=256 and r=16. Finally, we redistribute the weights of the different channels and prepare the data for the CSP-SPP.

The design of CSP-SPP aims to greatly enhance the feature extraction ability of the proposed model. The network is still divided into two groups shown in Fig. 5. Path A does not perform any image processing, whereas Path B uses SPP-Net and diverse convolutions to perform more comprehensive feature extraction. Eventually, we concatenate the two paths together, attaining a fusion of different information.

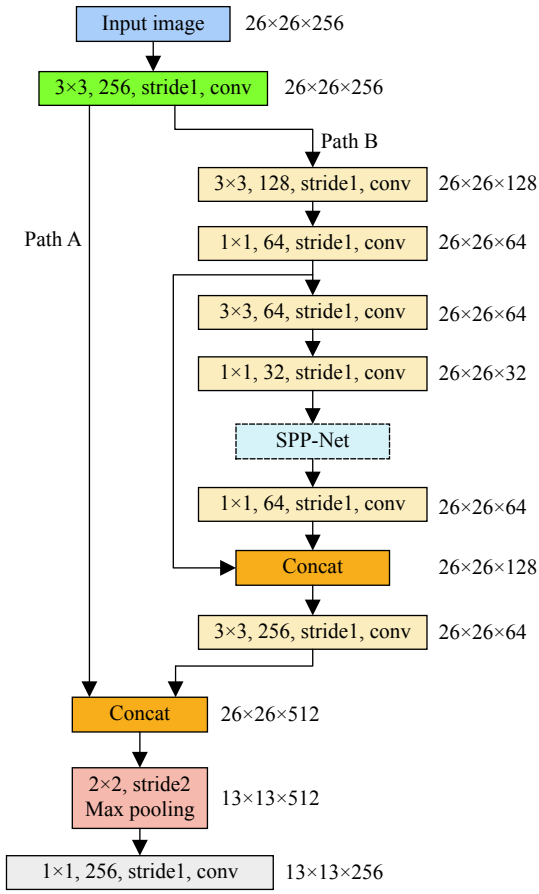


Fig. 5 Structure of CSP-SPP

To reduce the computational load, a 1×1 convolution is used to reduce the feature dimension before the data is input to the SPP-Net. As shown in Fig. 6, SPP-Net performs feature extraction from various receptive fields, and our results show that this network element is able to improve the detection accuracy effectively.

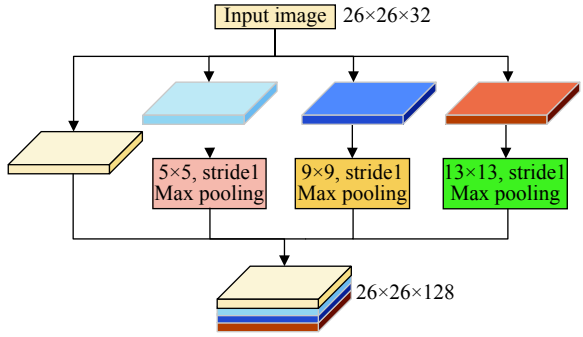


Fig. 6 Structure of SPP-Net

3.3 Accurate detection module

Object-detection networks generally have a two-part structure, involving feature extraction and object detection. LSM and EFM are first applied to achieve lightweight and efficient feature extraction. ADM is a multi-scale detection module designed for a higher detection accuracy. Considering that many target objects at low altitudes are small, ADM is added as a scale for detecting small flying objects, particularly compared with YOLOv4-tiny. Moreover, a lightweight network block (LNB) is presented before the final detection at each scale, as shown in Fig. 7. The block effectively improves the detection performance of the model with only a relatively insignificant increase in the amount of computation required.

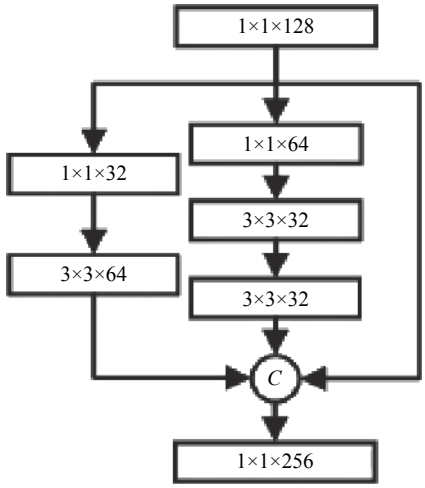


Fig. 7 LNB

To enable the model to perform more accurate detection, we adopt k-means clustering to redefine the size of anchors, which can enable the anchors to encode more representative prior information and perform more accurate prediction after regression. An example of the initialized centers and sizes of the bounding boxes is shown in Fig. 8.

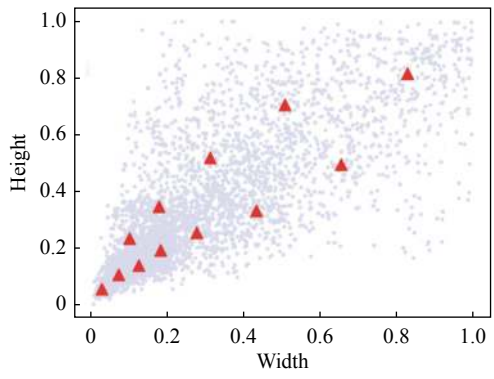


Fig. 8 Bounding and anchor distribution

Then, we iterate the cluster centers as follows:

$$d(\text{box},\text{centroid}) = 1 - \text{IoU}(\text{box},\text{centroid}). \tag{4}$$

In this formula, IoU means intersection over union of the truth boxes and the predicted boxes. The box and the centroid represent the sizes of the bounding boxes and the center of each cluster, respectively.

As shown in Table 1, three different sizes of feature maps are used, including 13 × 13, 26 × 26, and 52 × 52 maps. Large anchor boxes (155 × 128), (196 × 237), and (320 × 321) are used in the feature map with a size of 13 × 13 and 64 × 64 receptive fields to detect large objects. Medium-sized anchor boxes (55 × 126), (82 × 86), and (98 × 186) are applied to the 26 × 26 feature map with a 16 × 16 receptive field to detect objects of a medium volume. In the largest 52 × 52 feature map with an 8 × 8 receptive field, small anchor boxes (13 ×26), (33 × 48), and (56 × 64) are used to detect small objects.

Table 1 Specific size and distribution of anchor boxes

Feature-map size	Receptive field size	Anchor box size
13 × 13	32 × 32	320, 321, 196, 237, 155, 128
26 × 26	16 × 16	98, 186, 82, 86, 55, 126
52 × 52	8 × 8	56, 64, 33, 48, 13, 26

LSL-Net uses regression to optimize the detection problem, and the loss function contains three parts, which can be expressed as follows:

$$\text{loss} = \text{loss}_1 + \text{loss}_2 + \text{loss}_3 \tag{5}$$

where loss_1, loss_2 and loss_3 are the confidence, classification, and bounding box regression loss functions, respectively. The confidence loss function is expressed as follows:

$$\text{loss}_1 = -\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[C_i^j \log\left(\widehat{C}_i^j\right) + \left(1-\widehat{C}_i^j\right)\log\left(1-C_i^j\right)\right] - \lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[C_i^j \log\left(\widehat{C}_i^j\right) + \left(1-\widehat{C}_i^j\right)\log\left(1-C_i^j\right)\right] \tag{6}$$

where S^2 is the value of the input image grid numbers, B is the number of bounding boxes in a grid, and I_{ij}^{obj} indicates whether the object appears in the jth bounding box of the ith grid. If there is an object in the grid, its value is 1; otherwise I_{ij}^{obj}=0. C_i^j and $\widehat{C}_i^j$ are the confidence scores of the ground truth and predicted boxes, respectively, while λ_{noobj} is a weight parameter.

The classification loss function is given as follows:

$$\text{loss}_2 = -\sum_{i}^{S^2}\sum_{j}^{B} I_{ij}^{obj}\cdot$$

$$\sum_{c=1}^{C}\left[\hat{p}_i^j(c)\log\left(p_i^j(c)\right)-\left(1-\hat{p}_i^j(c)\right)\log\left(1-p_i^j(c)\right)\right]. \quad (7)$$

In this formula, $p_i^j(c)$ and $\hat{p}_i^j(c)$ are the prediction probability and the truth probability belonging to the jth bounding box of the ith grid, respectively.

$$\text{loss}_3 = 1-\text{IoU}+\frac{\rho^2(b,b^{gt})}{c^2}+\alpha\beta \quad (8)$$

where b and b^{gt} represent the center points of the bounding box and the ground truth, respectively, $\rho^2(b,b^{gt})$ is the Euclidean distance between the two center points belonging to the bounding box and the ground truth, and c indicates the diagonal distance of the minimum closure area that includes them, α is a tradeoff parameter, and β reflects the consistency of the length-width ratios. The formulas for calculating α and β are as follows:

$$\alpha=\frac{\beta}{1-\text{IoU}+\beta}, \quad (9)$$

$$\beta=\frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}}-\arctan\frac{w}{h}\right)^2, \quad (10)$$

where w and h represent the width and the height of the bounding box, respectively, and w^{gt} and h^{gt} represent those of the ground truth box.

3.4 LSL-Net overview

We propose an object detection network called LSL-Net consisting of an LSM, an EFM, and an ADM, as shown in Fig. 9. The model makes only one forward calculation to generate sufficient anchor boxes, similar to the YOLO algorithm. After assessing the confidence of each category, the eventual outcome is determined by using non-maximum suppression (NMS). The proposed method uses various effective image-enhancement techniques and fixes the resolution of images of different sizes. First, images are stably downsampled by the LSM (Fig. 1), enhancing the feature extraction performance without introducing additional calculations. The EFM (Fig. 2) then uses group convolution, CSPNet (Fig. 3), the attention mechanism (Fig. 4) and CSP-SPP (Fig. 5) to construct a parallel stacked identical topology with a strong feature extraction ability. The LSM and EFM achieve downsampling and extraction of high-quality image features. In the detection module, the detection scale, particularly for small objects, is added to the ADM. Moreover, the LNB (Fig. 7) and k-means clustering (Fig. 8) are adopted to achieve high-precision real-time detection.

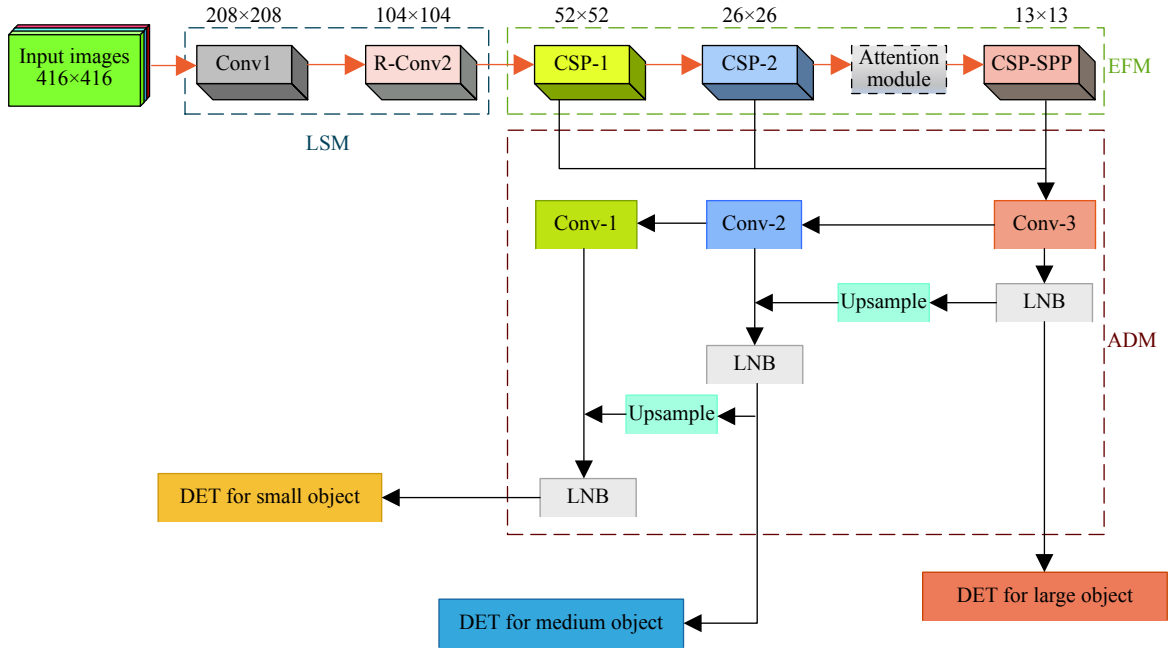


Fig. 9 Architecture of the LSL-Net

4. Tests and results

We build a new dataset by sampling low-altitude scenes to evaluate the detection capability of our model. The experiments are performed using a Ubuntu16.04 system, an Intel® CoreTM I7-6950X CPU, and four NVIDIA Ge-

Force GTX 1080Ti graphics cards with 11 GB of memory each. A graphics processing unit (GTX1080Ti) is used for training and testing.

4.1 Data preparation

Currently, numerous public datasets such as the Pascal visual object classes (VOC) and common objects in context (COCO) datasets are available for object detection in general scenes, but are not effective in low-altitude environments. We create a new dataset called LA2021, containing different types of weather at low altitudes. We annotate 6 904 images, including various possible low-altitude flying objects. The collected images are divided into three categories, including kites, birds, and UAVs. Birds and kites are relatively similar objects to UAVs at low altitudes. The purpose of identifying them is to perform UAV targeting tasks better. Fig. 10 displays a portion of the low-attitude dataset.

Fig. 10 Images of low-attitude environments

4.2 Results for LA2021

We use the stochastic gradient descent method to optimize the model. The batch size in the training process is 64, and the momentum is 0.9. The maximum number of batches is 1.2×10^5. For the first 10 000 iterations, we set the learning rate to 0.000 1 and the weight decay to 0.000 5. Then, at 1.8×10^4 and 2×10^4 iterations, we set the learning rate to 0.000 5 (a reduction by a factor of 0.1). Moreover, the mosaic data-enhancement method is applied to our model to enhance and extend the dataset.

4.2.1 Comparison of various frameworks on LA2021

We enumerate the detection results of LSL-Net compared to those of other models used in the Video Object Tracking Realtime Challenge 2019 (VOT-RT2019). To intuitively compare these several methods, we use the mAP and FPS as model performance evaluation indices. Table 2 presents the detection results of widely applied models.

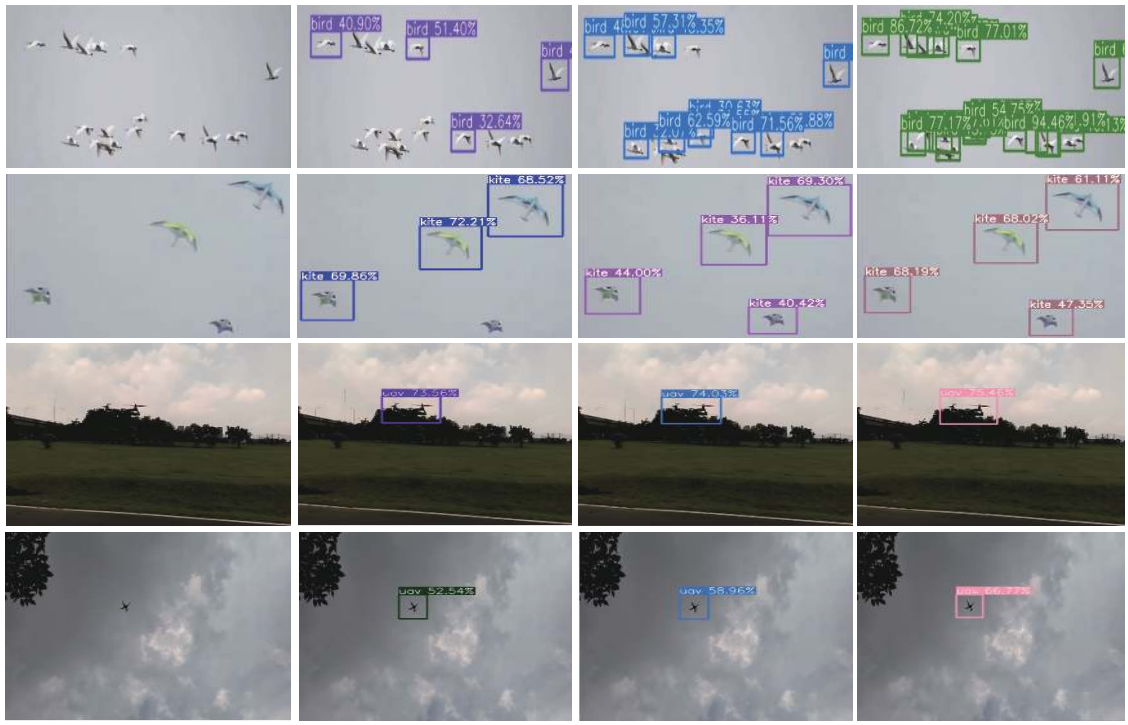
Table 2 Comparison of test results for LA2021

Component	Faster	SSD	Mbv2-SSD	YOLOv3	YOLOv3-tiny	FOCS	Center Net	YOLOv4	YOLOv4-tiny	LSL-Net
Bird	0.897 6	0.835 4	0.711 5	0.941 0	0.803 9	0.937 6	0.934 4	0.962 8	0.836 0	0.905 4
Kite	0.785 1	0.729 2	0.679 4	0.861 7	0.685 6	0.880 7	0.867 1	0.916 7	0.757 4	0.871 4
UAV	0.925 6	0.912 5	0.816 5	0.951 3	0.894 0	0.967 5	0.957 0	0.981 3	0.934 4	0.952 3
mAP	0.869 5	0.825 7	0.735 8	0.918 0	0.794 5	0.928 6	0.919 5	0.953 6	0.842 6	0.909 7
FPS	12	43	81	77	334	42	89	49	287	147

The experimental results in Table 2 indicate that LSL-Net strikes the best compromise between detection speed and accuracy, in contrast to other methods. Its detection accuracy and real-time performance are better than those of Faster-RCNN, SSD, and MobileNetv2-SSD, enabling it to identify low-altitude objects efficiently and accurately. Moreover, it includes a powerful UAV detection capability. The UAV detection accuracy of LSL-Net is 2.67% and 3.89% higher than that of Faster-RCNN and of SSD, respectively. Although its detection accuracy is lower than that of YOLOv3, YOLOv4, CenterNet, and FOCS, the detection speed of the proposed network is ap-

proximately double, three times, 1.5 times, and 3.5 times that of the other network models, respectively. The detection speed of LSL-Net is slower than that of YOLOv3-tiny and YOLOv4-tiny. However, the detection accuracy of LSL-Net is 11.52% and 6.71% higher than that of YOLOv3-tiny and of YOLOv4-tiny, respectively, and its real-time performance is sufficient to the industrial requirements. The results indicate that LSL-Net balances detection speed (147 FPS) and accuracy (90.97%) well and is the most suitable model for the detection of low-altitude flying objects.

We present the detection results of different lightweight models on the same dataset of low-attitude images in Fig. 11, showing that LSL-Net has the best performance in all the object categories. Each column presents the original image and the detection results of YOLOv3-tiny, YOLOv4-tiny, and LSL-Net from left to right. The first row of images shows a multitude of birds flying at low attitudes. YOLOv3-tiny misses more than half of the objects. The missed detection rate for birds of YOLOv4-tiny is obviously reduced, but there are still some omissions. However, LSL-Net is able to detect all the objects effectively, even if they were crowded together, and the confidences are superior to the previous two models. In the second row, the scale of the object changes slightly. YOLOv3-tiny misses a kite, whereas both YOLOv4-tiny and LSL-Net demonstrate better detection performance. However, the detection accuracy of LSL-Net is higher owing to its powerful detection capability. In the third row, there is only one UAV in the picture. However, the entire UAV is not visible in the detection frame owing to background interference. LSL-Net effectively identifies the UAV with accuracies 1.9% and 1.43% higher than those of YOLOv3-tiny and YOLOv4-tiny, respectively, which highlights the object detection ability of the model in complex background environments. In the images of the last row, there is a small UAV in the dim light. As indicated by the test results, LSL-Net is still able to perform efficient and accurate detection, and the test result is optimal. The results indicate that LSL-Net has a good robustness and can accurately determine the locations of objects, ensuring a high detection accuracy at low attitudes, particularly for small targets.



(a) Origin images (b) YOLOv3-tiny (c) YOLOv4-tiny (d) LSL-Net (ours)

Fig. 11 Detection results of different models

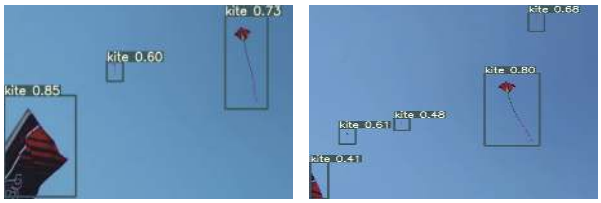
4.2.2 Robustness tests in different environments

Fig. 12 shows typical detection results for LSL-Net under various low-attitude scenarios. The network could detect low-altitude flying objects precisely, ensuring the successful completion of anti-UAV missions even for small targets in low-light environments (as shown in Fig. 12(a)), exhibiting the strong multi-object detection

ability of the proposed model at low attitudes (Fig. 12(b)). Although low flying objects tend to be very small, the low-attitude detection ability of LSL-Net remains effective (Fig. 12(c)). Experimental results on the robustness of the proposed model show that LSL-Net can detect different object locations accurately in various scenes with high robustness, satisfying the requirements for anti-UAV tasks, and ensuring the safety and security of public property.

(a) Small object in low-light environment

(b) Multi-objective of low-attitude



(c) Minimal objects of multiple flyings

Fig. 12 Robustness test results

4.2.3 Ablation experiments

To evaluate the efficiencies of the LSM, EFM, and ADM, we design and test three diverse models. In a model called EA-Net, the LSM is omitted. For a mode called LA-Net, the EFM is omitted. For the final LE-Net, the ADM is omitted. Omitting the abovementioned modules, the structure of the original network (YOLOv4-tiny) is replaced. The experimental results are in Table 3. First, we add an LSM to EA-Net. The mAP is increased by 0.35% when the detection speed is increased. It is thus obvious that the LSM can improve the precision of low-flying object detection and reduce the loss of the original input image information. Then, the original feature extraction network in LA-Net is replaced with an EFM. As a result, we enhance the detection accuracy by 2.02% without increasing the computational load. These results indicate that the EFM implemented by various optimizing convolutions and attention mechanisms allows the extraction of more feature information and effectively improves the detection accuracy of the proposed model. Finally, we add the ADM to LE-Net. The detection accuracy of LE-Net is relatively low compared to the other methods, and is then significantly improved with the addition of the ADM, particularly for small objects. The experimental results indicate that the EAM significantly improves object detection performance. Although YOLOv4-tiny has the highest detection speed among the models tested, its poor detection accuracy cannot meet the requirements in complex low-altitude environments. In contrast, LSL-Net operates with an excellent balance between detection speed and accuracy, and can satisfy the application requirements for the detection of low-altitude flying objects.

Table 3 Effects of various design choices

Component	YOLOv4-tiny	EA-Net	LA-Net	LE-Net	LSL-Net
LSM	—	—	√	√	√
EFM	—	√	—	√	√
ADM	—	√	√	—	√
Bird	0.836 0	0.900 5	0.871 7	0.831 2	0.905 4
Kite	0.757 4	0.868 3	0.856 2	0.806 2	0.871 4
UAV	0.934 4	0.949 8	0.940 6	0.931 5	0.952 3
mAP	0.842 6	0.906 2	0.889 5	0.856 3	0.909 7
FPS	287	141	169	253	147

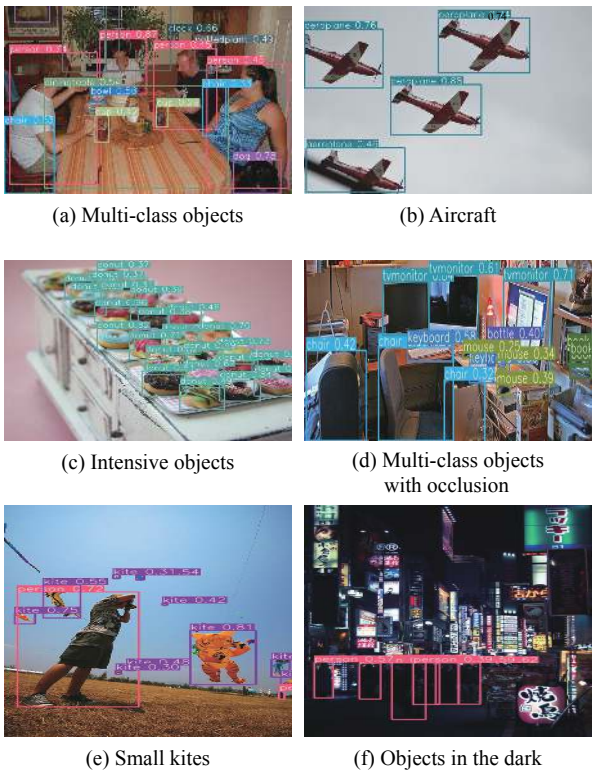
4.3 Results on MS COCO dataset

To verify the migration ability of LSL-Net, we test our model on the MS COCO [42] (test-dev2017) dataset, and a comparison of the test results with other state-of-the-art models is shown in Table 4. In contrast to the UAV dataset, the MS COCO dataset includes a total of 80 classes of objects with larger object scales and more small objects, which means more complex environments and more challenging detection tasks. It is proven experi-

mentally that LSL-Net realizes the best compromise between the detection speed and the accuracy of the algorithm. The detection speed and the accuracy of LSL-Net are very significantly higher than those of YOLOv3. Although the detection accuracy of LSL-Net is reduced by 3.2% compared with YOLOv4, the detection speed is almost four times that of YOLOv4, enabling better object detection in real time. The results show that LSL-Net has good generalization and can be applied to complex scenes. Partial detection results of LSL-Net on the MS COCO dataset are shown in Fig. 13. The results indicate that LSL-Net can be effectively applied in a detection scene with multi-class objects (Fig. 13(a) and Fig. 13(d)). Even in intensive object detection environments (Fig. 13(c)) and dim environments (Fig. 13(f)), LSL-Net is still able to detect objects accurately. In particular, there is no missed detection of aircraft (Fig. 13(b)) by using LSL-Net, and even small kites could be accurately identified (Fig. 13(e)).

Table 4 Results on MS COCO

Method	Size	mAP/%	FPS
Faster R-CNN[21]	—	39.8	9
SSD[22]	300×300	25.1	43
SSD[22]	512×512	28.8	22
YOLOv3-SPP[25]	608×608	36.2	20
YOLOv4[26]	608×608	43.5	33
CenterNet[27]	—	41.6	28
FCOS[28]	—	44.7	—
LSL-Net(ours)	416×416	38.4	135
LSL-Net(ours)	512×512	39.1	126
LSL-Net(ours)	608×608	40.3	118



(a) Multi-class objects (b) Aircraft

(c) Intensive objects (d) Multi-class objects with occlusion

(e) Small kites (f) Objects in the dark

Fig. 13 MS COCO results of LSL-Net

5. Conclusions

We propose LSL-Net to perform high-precision detection of low-altitude flying objects in real time to provide information as guidance to suppress black flight of UAVs. The model comprises three simple and efficient modules, including LSM, EFM, and ADM. The LSM reduces the image input size and the loss of low-level feature information. The EFM improves the feature extraction ability of the model by using an attention mechanism and CSPNet, and the ADM increases the detection accuracy, especially for small objects. A dataset of low-altitude flying objects containing multi-class and multiscale objects is constructed to evaluate the performance of the proposed network. In an experiment, LSL-Net achieves an mAP of 90.97% and a detection speed of 147 FPS on an NVIDIA GTX1080Ti (6.71% higher than YOLOv4-tiny and 98 FPS faster than YOLOv4, respectively). The results of numerous experiments indicate that LSL-Net, which has a good robustness and an excellent generalization ability, can effectively perform detection of different weather conditions and satisfy the requirements of low-altitude flying object detection for anti-UAV missions. Moreover, experiments on MS COCO indicate that LSL-Net is also suitable for object detection in other complex scenes. In the future, we will design an im-

proved model that can be adapted to devices such as embedded mobile terminals, and further enrich the dataset, adding object categories in more complex backgrounds to make it more representative.

References

[1] ZHANG Y, SI G Y, WANG Y Z. Simulation of unmanned aerial vehicle swarm electromagnetic operation concept. Systems Engineering and Electronics, 2020, 42(7): 1510–1518. (in Chinese)

[2] MA Q, ZHU B, CHENG Z D, et al. Detection and recognition method of fast low-altitude unmanned aerial vehicle based on dual channel. Acta Optica Sinica, 2019, 39(12): 105–115.

[3] DONG Q, ZOU Q H. Visual UAV detection method with online feature classification. Proc. of the IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference, 2017: 429–432.

[4] HU Y Y, WU X J, ZHENG G D, et al. Object detection of UAV for anti-UAV based on improved YOLO v3. Chinese Control Conference, 2019: 8386–8390. (in Chinese)

[5] ZHANG H, CAO C H, XU L W, et al. A UAV detection algorithm based on an artificial neural network. IEEE Access, 2018, 6(1): 24720–24728.

[6] ZHAO W C, ZHANG Q, LI H, et al. Low-altitude UAV detection method based on one-staged detection framework. Proc. of the 2nd International Conference on Advances in Computer Technology, Information Science and Communications, 2020: 112–117.

[7] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.

[8] LOW D G. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110.

[9] BAY H, TUYTELAARS T, GOOL L V. SURF: speeded up robust features. Computer Vision & Image Understanding, 2006, 110(3): 404–417.

[10] FREUND Y. Experiments with a new boosting algorithm. Proc. of the 13th International Conference on Machine Learning, 1996: 148–156.

[11] WU S Q, NAGAHASHI H. Parameterized AdaBoost: introducing a parameter to speed up the training of real AdaBoost. IEEE Signal Processing Letters, 2014, 21(6): 687–691.

[12] WANG X H, ZHANG C, LI C, et al. Unmanned aerial vehicles target detection based on bio-inspired visual attention. Aeronautical Science & Technology, 2015, 26(11): 76–82.

[13] OMID-ZOHOOR A, YOUNG C, TA D, et al. Toward always-on mobile object detection: energy versus performance tradeoffs for embedded hog feature extraction. IEEE Trans. on Circuits and Systems for Video Technology, 2018, 28(5): 1102–1115.

[14] LIU Y, DU H J, YUE Z H, et al. Unmanned aerial vehicle detection method based on random forest. Computer Engineering and Applications, 2019, 55(7): 162–167.

[15] LI Y Z, XIE P C, TANG Z S, et al. SVM-based sea-surface small target detection: a false-alarm-rate-controllable approach. IEEE Geoscience and Remote Sensing Letters, 2019, 16(8): 1225–1229.

[16] BAZI Y, MELGANI F. Convolutional SVM networks for object detection in UAV imagery. IEEE Trans. on Geoscience and Remote Sensing, 2018, 56(6): 3107–3118.

[17] ZEILER M D, FERGUS R. Visualizing and understanding convolutional neural networks. Proc. of the European Conference on Computer Vision, 2014: 818–833.

[18] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580–587.

[19] GIRSHICK R. Fast R-CNN. Proc. of the IEEE International Conference on Computer Vision, 2015: 1440–1448.

[20] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149.

[21] WANG J Q, CHEN K, YANG S, et al. Region proposal by guided anchoring. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 2965–2974.

[22] ANGUELOV D, ERHAN D, SZEGEDY C, et al. SSD: single shot multibox detector. Proc. of the European Conference on the Computer Vision, 2016: 21–37.

[23] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779–788.

[24] REDOMN J, FARHADI A. YOLO9000: better, faster, stronger. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6517–6525.

[25] REDOMN J, FARHADI A. YOLOv3: an incremental improvement. https://arXiv.arg/abs/1804.02767.

[26] BOCHKOVSKIY A, WANG C Y, MARK LIAO H Y. YOLOv4: optimal speed and accuracy of object detection. https://arXiv.arg/abs/2004.10934.

[27] ZHOU X, WANG D, KRAHENBUHL P. Objects as points. https://arXiv.arg/abs/1904.07850.

[28] TIAN Z, SHEN C H, CHEN H, et al. FCOS: fully convolutional one-stage object detection. Proc. of the IEEE/CVF International Conference on Computer Vision, 2019: 9626–9635.

[29] LI Q Z, XIONG R, WANG R P, et al. Research on real-time UAV recognition method based on SSD algorithm. Ship Electronic Engineering, 2019, 39(5): 30–35. (in Chinese)

[30] KRIZHEVSKY A, SUTSKEVER I, GEOFFREY E H. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90.

[31] RANGA A, TYAGI A, BERG A C, et al. DSSD: deconvolutional single shot detector. https://arXiv.arg/abs/1701.06659.

[32] ZENG J X, XIONG J L, FU X, et al. ReFPN-FCOS: one-stage object detection for feature learning and accurate localization. IEEE Access, 2020, 8: 225052–225063.

[33] FAN J X, LI D W, WANG H L, et al. UAV low altitude flight threat perception based on improved SSD and KCF. Proc. of the IEEE 15th International Conference on Control and Automation, 2019: 121–132.

[34] FAN S Q, ZHU F H, CHEN S C, et al. FII-CenterNet: an anchor-free detector with foreground attention for traffic object detection. IEEE Trans. on Vehicular Technology, 2021, 70(1): 121–132.

[35] MA Q, ZHU B, ZHANG H W, et al. Low-altitude UAV detection and recognition method based on optimized YOLOv3. Laser & Optoelectronics Progress, 2019, 56(20): 279–286.

[36] CUI Y P, WANG Y H, HU J W. Detection method for a dynamic small target using the improved YOLOv3. Journal of Xidian University, 2020, 47(3): 1–7. (in Chinese)
[37] ARTUUR D, VASSILVITSKII S. k-means++: the advantages of careful seeding. Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 2007: 1027 – 1035.
[38] WEI X K, WEI D H, SUO D, et al. Multi-object defect identification for railway track line based on image processing and improved YOLOv3 model. IEEE Access, 2020, 8: 61973–61988.
[39] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 1571–1580.
[40] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. computer vision and pattern recognition. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.
[41] XIE S N, GIRSHICK R, DOLLAR P, et al. Aggregated residual transformations for deep neural networks. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5987–5995.
[42] TSUNG Y L, MICHAEL M, BELONGIE S, et al. Microsoft COCO: common objects in context. Proc. of the European Conference on Computer Vision, 2014: 740–755.

Biographies

YE Tao was born in 1987. He received his B.S. degree in measurement and control technology and instrumentation from China University of Mining and Technology, Xuzhou, China, in 2009, M.S. degree in mechanical and electronic engineering from China University of Mining and Technology-Beijing, Beijing, China, in 2012, and Ph.D. degree in measurement technology and instruments from the Key Laboratory of Precision Opto-mechatronics Technology of Ministry of Education, Beihang University, Beijing, in 2016. He was an engineer with Beijing Institute of Remote Sensing and Equipment from 2016 to March 2019. He is currently a senior engineer with the School of Mechanical Electronics and Information Engineering, China University of Mining and Technology-Beijing, Beijing. His current research interests include deep learning and traffic detection.
E-mail: ayetao198715@163.com

ZHAO Zongyang was born in 1998. He received his B.S. degree in mechanical engineering from China University of Mining and Technology-Beijing, China, in 2020, where he is currently pursuing his M.S. degree with the Department of Mechanical Engineering, School of Mechanical Electronic and Information Engineering. His current research interests include deep learning and railway object detection.
E-mail: 303616426@qq.com

ZHANG Jun was born in 1998. He received his B.S. degree in mechanical engineering from China University of Mining and Technology-Beijing in 2020, Beijing, China. He is currently pursuing his M.S. degree in mechanical engineering at China University of Mining and Technology-Beijing. His research interests include artificial intelligence, deep learning, and multiple-object tracking.
E-mail: 973974045@qq.com

CHAI Xinghua was born in 1986. He received his B.S. degree from Hohai University in 2008, M.S. degree from Beijing Information Science and Technology University in 2013, and Ph.D. degree from Beihang University in 2017. He is working as a senior engineer with the 54th Research Institute of the China Electronics Technology Group Corporation. His research interests include machine vision, computer vision, and multi-agent systems.
E-mail: cxh88_88@163.com

ZHOU Fuqiang was born in 1972. He received his B.S., M.S., and Ph.D. degrees in instrument, measurement, and test technology from Tianjin University, China, in 1994, 1997, and 2000, respectively. He joined the School of Automation Science and Electrical Engineering, Beihang University, China, as a post-doctoral research fellow, in 2000. He is currently a professor with the School of Instrumentation and Opto-electronics Engineering, Beihang University. His research interests include precision vision measurement, 3D vision sensors, image recognition, and optical metrology.
E-mail: zfq@buaa.edu.cn