Computers, Materials & Continua
DOI: 10.32604/cmc.2023.034876
Article

Tech Science Press

RT-YOLO: A Residual Feature Fusion Triple Attention Network for Aerial Image Target Detection

Pan Zhang, Hongwei Deng* and Zhong Chen

College of Computer Science and Technology, Hengyang Normal University, Hengyang, 421002, China
*Corresponding Author: Hongwei Deng. Email: dhwwhd@163.com

Received: 30 July 2022; Accepted: 14 December 2022

Abstract: In recent years, target detection of aerial images of unmanned aerial vehicle (UAV) has become one of the hottest topics. However, target detection of UAV aerial images often presents false detection and missed detection. We proposed a modified you only look once (YOLO) model to improve the problems arising in object detection in UAV aerial images: (1) A new residual structure is designed to improve the ability to extract features by enhancing the fusion of the inner features of the single layer. At the same time, triplet attention module is added to strengthen the connection between space and channel and better retain important feature information. (2) The feature information is enriched by improving the multi-scale feature pyramid structure and strengthening the feature fusion at different scales. (3) A new loss function is created and the diagonal penalty term of the anchor frame is introduced to improve the speed of training and the accuracy of reasoning. The proposed model is called residual feature fusion triple attention YOLO (RT-YOLO). Experiments showed that the mean average precision (mAP) of RT-YOLO is increased from 57.2% to 60.8% on the vehicle detection in aerial image (VEDAI) dataset, and the mAP is also increased by 1.7% on the remote sensing object detection (RSOD) dataset. The results show that the RT-YOLO outperforms other mainstream models in UAV aerial image object detection.

Keywords: Attention mechanism; small target detection; YOLOv5s; RT-YOLO

1 Introduction

With the continuous development of artificial intelligence technology and computer hardware conditions, artificial intelligence has made a series of research progress in many fields, and human development has gradually entered the era of intelligence. Driven by scientific and technological innovation, all walks of life in society are exploring how to introduce artificial intelligence into their respective fields. In the unmanned aerial vehicle (UAV) field, UAV aerial photography offers a wide range of potential applications in military and public domains, including military reconnaissance, emergency disaster relief, land surveying, mapping, and agricultural plant protection [1].

In the future process of intelligent urban development, object detection involves identifying and classifying target positions from aerial photographs. It has become one of the most important tasks in the field of UAV aerial photography. However, the complex background of UAV aerial images, the uneven distribution of targets, a large number of small targets and overlapping targets. The detection effect of the existing general algorithms is not very good. Therefore, researchers have conducted many studies on UAV aerial image detection methods.

Current UAV aerial object detection techniques are based on convolutional neural network (CNN) [2] and are classified into two categories: a two-stage approach based on candidate regions and a single-stage strategy based on regression. The two-stage technique searches for the region of interest using region proposal network (RPN) and then generates the category and location information for each region. R-CNN [3], Fast R-CNN [4], Faster R-CNN [5], and Mask R-CNN [6] are examples of algorithms. Object detection is treated as a regression problem in the one-stage algorithm. The output layer collects the target's position and category information instantly after putting the target to gets detected. The YOLO series [7–11] and the single shot multibox detector (SSD) series [12,13] are two examples of representative algorithms. Moreover, Bera et al. [14] carried out a detailed performance and analysis of four CNN models: 1D CNN, 2D CNN, 3D CNN, and feature fusion based on CNN (FFCNN).

Many small object detection algorithms based on deep learning have been proposed in recent years by researchers. Based on general detectors, they have tried to make improvements from different directions such as accuracy and speed, including data enhancement, multi-scale learning, context learning, and combined with generative adversarial network (GAN) and other methods. Sommer et al. [15] used Fast R-CNN and Faster R-CNN for vehicle detection in aerial images to accommodate small target detection by adjusting the size of the anchor frame and the resolution of the feature map. A simple and successful approach to ratio matching was described in the document [16]. Images are scaled and spliced throughout the training process, and large-size targets in the dataset are transformed into medium-size targets and medium-size targets into small-size targets. Increase the number and quality of small-scale goals. Yang et al. [17] introduced the attention mechanism into target detection. Supervised multi-dimensional attention network (MDA-NET) is used to highlight the target features and weaken the background features. Ibrahim et al. proposed an adaptive dynamic particle swarm algorithm [18]. In combination with guided whale optimization algorithm (WOA), the prediction performance of the algorithm is improved by enhancing the parameters of the long short-term memory (LSTM) classification method. Rao et al. used the newly designed rectified linear unit (ReLU) to propose a new model [19], which inserts a ReLU layer before the convolution layer. This structure can more smoothly transfer semantic information from the shallow layer to the deep layer. It prevented network degradation and improved the performance of deep networks. Yang et al. proposed the semi-supervised attention (SSA) model [20], which has a semi-supervised attention structure for different small target images. Using unlabeled data in the data can help reduce the change of the same category and achieve more distinguishing feature extraction. Singh et al. introduced scaling normalization of image pyramid (SNIP) [21], a multi-scale training strategy that trains on each scale of the pyramid and efficiently employs all of the training data, despite the detection effect of small targets. Although there has been a tremendous improvement, the speed has slowed. High-resolution network (HRNet) [22] was published. It has achieved considerable progress in inaccuracy by using a parallel structure to fuse feature maps of many scales to generate more resilient multi-scale feature information. Xu et al. [23] proposed a novel relational graph attention network that incorporates edge attributes. Considers the edge attributes by using top-k attention mechanisms to learn hidden semantic contextual, improved network performance. Chen et al. [24] presented an improved YOLOv4 algorithm, which increases the dimension of the effective feature layer of the backbone network. It introduces the cross stage partial (CSP) structure into path aggregation network (PANet). The computational complexity of the model is reduced, and the polymerization efficiency of effective features at different scales is improved. Article [25] proposed a multi-scale symbolic method, which combines symbolization and multi-scale technology with compression to enhance the ability of feature extraction. Su et al. [26] combined UAV sensing, multispectral imaging, vegetation segmentation, and u-net to design a spectrum-based classifier and conduct a systematic evaluation to improve performance in UAV visual perception. Paper [27] proposed multi-objective artificial hummingbird algorithm (MOAHA). A non-dominated sorting strategy is merged with MOAHA to construct a solution update mechanism, which effectively refines Pareto optimal solutions for improving the convergence of the algorithm.

To solve the problem that the features of low-resolution small targets cannot be detected, some scholars have combined the generation of confrontation networks with detection models. They proposed methods such as Perceptual GAN [28], SOD-MTGAN [29], and CGAN [30]. The complexity of generating an adversarial network is too high to meet the needs of UAV aerial image target detection. Therefore, some researchers advocated combining lightweight network models such as the MobileNet [31–33] series, ShuffleNet [34,35] series, and others to be used in real-world applications.

Currently, the performance of commonly used object detection algorithms in detecting aerial picture targets is limited. There are mainly the following problems in aerial image object detection:

(1) In a UAV aerial image, the size range of the target is vast, the proportion of the small target in the image is very small, and the resolution provided is limited, which is difficult to detect.
(2) There are identical items in the dense zone of targets in UAV aerial pictures, resulting in a higher incidence of missing or false alarm detection. Furthermore, a significant amount of background noise information will weaken or obscure the target, making consistent and complete identification impossible.

Based on the aforementioned issue, we propose an improved YOLOv5 model named residual feature fusion triple attention YOLO (RT-YOLO). First, the newly designed residual module is used to enhance the utilization of single-layer internal features, while adding a triplet attention module to establish spatial and channel connections, and improve the ability of the backbone network to extract features. Then, the feature pyramid structure is improved to integrate the feature maps at various scales and reduce the loss of feature information. Finally, we proposed a new loss function and introduced a diagonal penalty term for the anchor frames, which improves training speed and inference accuracy. On the vehicle detection in aerial image (VEDAI) dataset [36] and remote sensing object detection (RSOD) [37] dataset, we compare RT-YOLO to other advanced object detection algorithms. Experimental results show that RT-YOLO is more suitable for object detection in UAV aerial images.

The rest of this paper is organized as follows. Section 2 introduces the related work, including the UAV aerial datasets, YOLOv5, and triplet attention module. Section 3 introduces the method set out in the present paper. Section 4 presents the details of the experiment, including the experimental datasets, experimental environment, and evaluation indicators. Section 5 introduces the relevant experiments and a discussion of the experimental results. Section 6 summarizes the work of this paper.

2 Related Work

2.1 Different Image Datasets

Because the aerial image of UAV is different from that of natural scene images, the target detection algorithm trained by using conventional image datasets is ineffective in the application of UAV scene tasks. Some researchers have proposed aviation image datasets for this problem, and the relevant image datasets are shown in Table 1.

Table 1: Comparison of different aerial image datasets

Dataset	Publish year	Number of pictures	Size of pictures	Number of class	Total of target
VEDAI	2015	1210	1024	9	3640
RSOD	2015	976	1044	4	6950
UCAS-AOD	2015	910	1280	2	6029
DOTA-v1.0	2017	2806	800–4000	15	188 282
DOTA-v1.5	2019	2806	800–4000	16	400 000
VisDrone	2019	10209	2000	10	89 777
Drone vehicle	2020	31064	840	5	441 642

Images in the VEDAI dataset are derived from Utah and are widely used for multi-variety vehicle detection tasks under aerial images. The dataset of object detection in aerial images (UCAS-AOD) dataset has simply two categories: aircraft and vehicle, for target detection of vehicles and aircraft under aerial images. The dataset for object detection in aerial images (DOTA) dataset is a large dataset, where images are acquired through different sensors and platforms. It includes target objects with different proportions, orientations, and shapes. The vision meets drone (VisDrone) dataset contains videos and images from various weather and light conditions. It can be utilized for four challenge tasks: UAV aerial image target detection, video target detection, single-target tracking, and multi-target tracking. The drone based vehicle detection (DroneVehicle) dataset contains red, green, and blue (RGB) images and infrared images of vehicle detection and vehicle counting tasks.

2.2 YOLOv5

YOLOv5 [11], which was released in 2020, is a regression-based target identification algorithm that comes in four versions: YOLOv5s, YOLOv5m, YOLOv51, and YOLOv5x. The network with the least depth and feature map width is YOLOv5s. This paper selects YOLOv5s with the minimum network depth and width for training to minimize processing costs and memory and make the network more lightweight, the exact structure is given in Fig. 1 below.

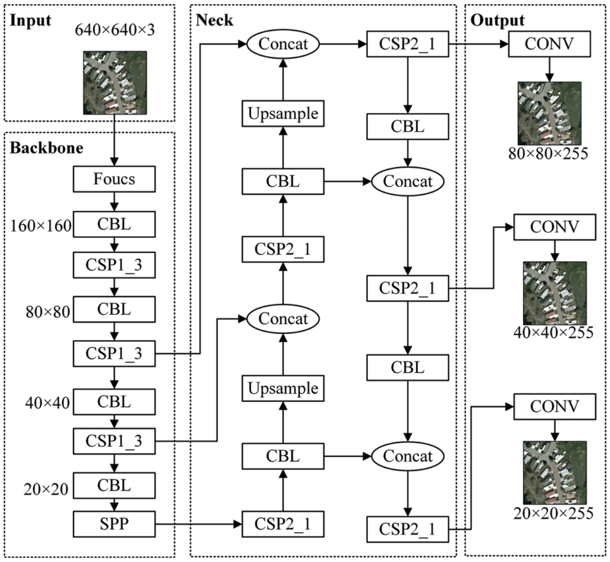
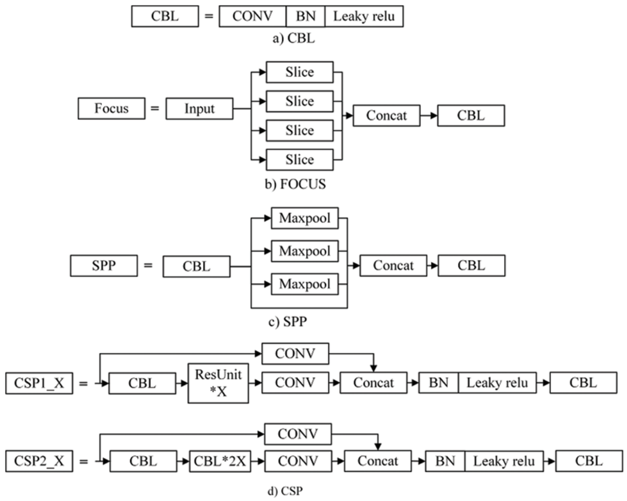


Figure 1: The structure of the YOLOv5 network

The input part of the YOLOv5 network is used for data preprocessing, the backbone section is used for feature extraction, the neck part is used for feature fusion, and the output part is used for object detection. CSPDarknet53 serves as the backbone network, and feature maps of various sizes are retrieved via repeated convolutions and merging. The trunk network generates four-layer feature maps, as shown in Fig. 1. When the input image is 640 × 640 pixels, the resulting feature maps are 160 × 160 pixels, 80 × 80 pixels, 40 × 40 pixels, and 20 × 20 pixels. The neck network uses the feature pyramid network (FPN) and path aggregation network (PAN). The FPN structure transmits semantic features from the top feature map to the bottom feature map, while the PAN structure transmits positioning features from the bottom feature map to the top feature map. Three feature fusion layers then fuse these different levels of feature maps. It can obtain more contextual information and generate three feature maps of various sizes. The output part recognizes and categorizes these feature maps of various sizes, which are 80 × 80 × 255, 40 × 40 × 255, and 20 × 20 × 255, where 255 denotes the number of channels, 80 × 80 × 255 features for identifying tiny things, and 20 × 20 × 255 features for detecting large objects, respectively.

We go over the basic module functionalities of yolov5 in depth to help you better grasp its architecture. The structure of each functional module is shown in Fig. 2 below. The convolution layer, batch normalization layer, and LeakyRelu activation functions (CBL) as shown in Fig. 2a. CBL is the smallest component of the YOLO network. The Focus module as shown in Fig. 2b, aids the backbone network in extracting features through slicing and connecting activities. The spatial pyramid pooling (SPP) module shown in Fig. 2c, performs feature fusion by maximizing pooling kernels of different sizes. Cross stage partial (CSP) structure are separated into two categories, as shown in Fig. 2d, with CSP1_X for the backbone network and CSP2_X for the neck network. CSP1_X is made up of X residual units, while CSP2_X is made up of CBL modules. Through cross-layer connection, the CSP structure decreases model complexity and speeds up reasoning speed.



a) CBL

b) FOCUS

c) SPP

d) CSP

Figure 2: Structure of each functional module of yolov5: (a) CBL; (b) Focus; (c) SPP; (d) CSP

2.3 Triplet Attention

Triplet attention [38] module to make full use of small target features. Through rotation operations and residual modifications, triplet attention builds the relationship between dimensions, which can improve the spatial and channel information qualities. Fig. 3 depicts the network structure.

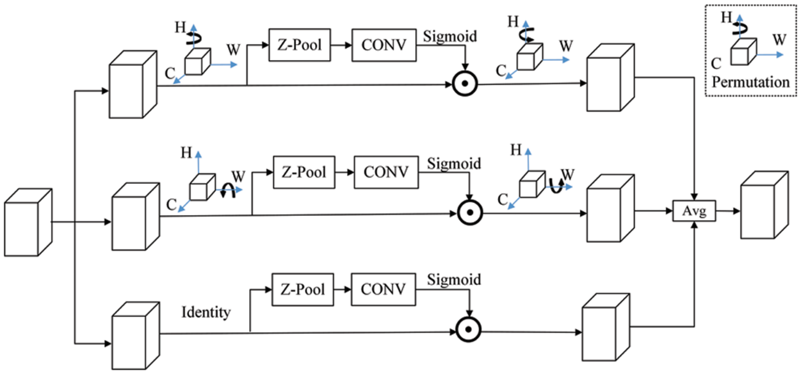


Figure 3: The module structure of triplet attention

The channel attention calculation branch, the channel C and space W dimension interactive capture branch, and the channel C and space H dimension interactive capture branch are the three branches of triplet attention. The indirect relationship between channels and weights can be avoided with this cross-channel interaction.

3 The Proposed Methods

We created a residual feature fusion triple attention network to boost the detection effect on the aerial small object detection challenge. Three improvements are made to the original yolov5 algorithm, as seen in Fig. 4.

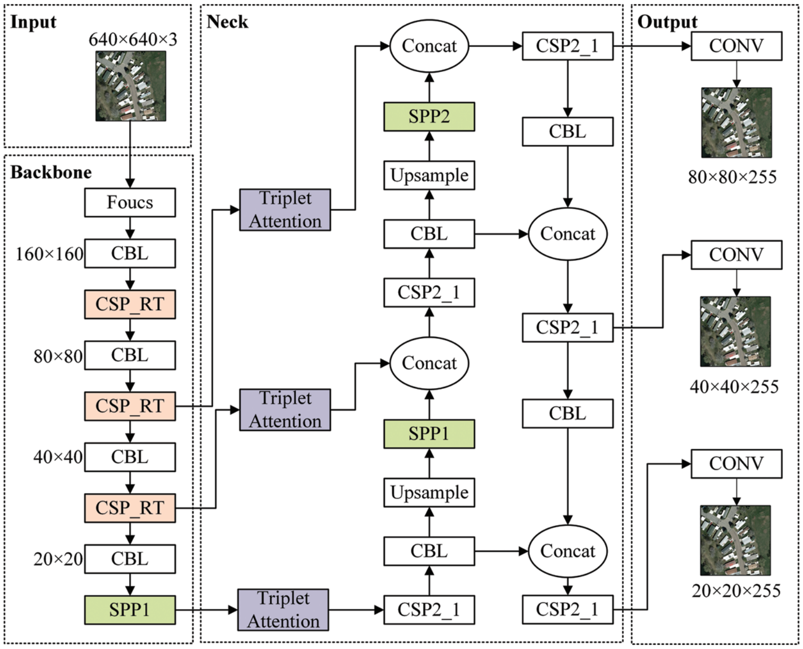


Figure 4: Network structure of RT-YOLO method

Small targets occupy the main distribution in the UAV images. The resolution provided is limited, which is difficult to detect. Firstly, we design the Res2T residual module and add triplet attention. It can express multi-scale features at a fine-grained level while capturing cross-dimensional interact information, which is more conducive to extracting Single-layer internal feature information. To simplify the backbone network's parameter complexity, the original CSP structure's short message channel convolution layer is removed, and the Res-unit in the long channel is substituted. We renamed it CSP_RT. Secondly, combining with the characteristics of the YOLO neck network, we improve the spatial pyramid pooling module by enhancing the algorithm's ability to extract small target features by increasing the fusing of several receptive fields and utilizing a lower maximum pooling to make the algorithm pay more attention to local information. Finally, we propose RIOU_Loss as the bounding box regression's loss function, which takes into account the aspect ratio, center point distance, and diagonal length. This effectively solves the situation where the prediction box is inside the target box and the size of the prediction box is the same, improving positioning accuracy and speeding up network convergence.

3.1 Feature Extraction Enhancement Module

Small targets make up a small percentage of UAV aerial photos, and the resolution offered is restricted, making feature extraction challenging. We designed the Res2T single-layer feature fusion module and improved the CSP feature extraction structure to overcome this problem.

3.1.1 Res2T

In the original YOLOv5 algorithm backbone network, the CSP feature extraction module takes advantage of the residual structure. Although the problem of gradient fading is mitigated as the method is deepened, the CSP structure continues to utilize the hierarchical multi-scale representation to represent features, leaving internal features of a single-layer underutilized, and the YOLOv5 algorithm gives equal attention to each channel feature. The algorithm's detection performance is restricted to some extent by this architecture. In response to these issues, we learned from the Res2Net module proposed by Gao et al. [39]. It combines with the triplet attention mechanism and designed the Res2T module. Fig. 5 depicts the specific structure.

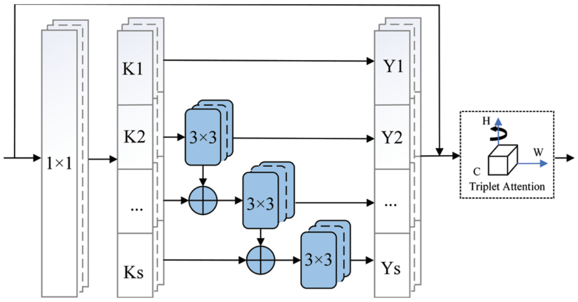


Figure 5: The structure of Res2T

In each Res2T structure, after a 1 × 1 convolutional layer, the input feature map is evenly divided into S sub-feature maps (S = 4 is selected in this article). The size of each sub-feature map is the same one, but the number of sub-feature map channels is 1/S of the number of input feature map channels. For each sub-feature map K_i, there is a corresponding 3 × 3 convolution, and the output is $Y_{i.}$ Each sub-feature map K_i is added with the output Y_{i-1} of K_{i-1}, which is used as the input of K_i corresponding to 3 × 3. To reduce the number of parameters, the 3 × 3 convolutional network of K_1 is omitted, which is specifically expressed as formula (1).

$$Y_i = \begin{cases} K_i, & i = 1 \\ (3 \times 3Conv)\left(K_i + Y_{i-1}\right), & x \geq 0 \end{cases} \tag{1}$$

As a control parameter, S is the number of input channels that divide multiple characteristic channels. The larger the S, the stronger the multi-scale capability of Res2T. Through different S, the output of different sizes of receptive fields can be obtained.

3.1.2 CSP_RT

In the original YOLOv5 algorithm, as the network convolution deepens, the feature information of the small target becomes weaker and weaker. It results in missed detection and false detection in aerial photography small target identification tasks. In the network optimization process, the backbone network's CSP structure eliminates repeating gradient information. Multiple convolution kernels, on the other hand, increase in the number of parameters as the network depth increases. To address the aforementioned issues, this article improves the CSP structure by deleting the convolutional layer on the original module's short branch, directly connecting the CSP module's input feature map with the output feature map of the long branch, and using the Res2T module to replace the residual connected unit of the CSP long branch. The improved feature extraction module is known as CSP_RT, and its structure is represented in Fig. 6.

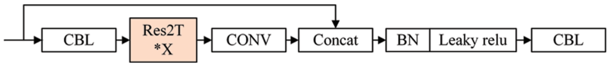


Figure 6: CSP_ RT module structure diagram

The CSP_RT structure, when compared to the original CSP structure, can extract shallow feature information more effectively, fuse crucial spatial and channel data in the feature map, and effectively increase the detection effect of small targets without increasing the number of parameters. Furthermore, multi-scale feature extraction improves the algorithm's semantic representation.

3.2 Feature Pyramid Module

A three-layer feature map detection design is used by the YOLOv5 algorithm. To detect targets of various sizes, feature maps sampled at 8 times, 16 times, and 32 times are employed as feature layers for the input picture scale of 640 × 640. To prevent the loss of tiny target information, each pixel should respond to the region corresponding to a small target in the image, corresponding to various output feature maps. As a result, through four groups of various maximum pooling layers, we modified the pooling module of the SPP spatial pyramid to improve the fusion of multiple receptive fields. To match the structure of YOLO output, the maximum pooling filter of 1 × 1, 3 × 3, 5 × 5, 7 × 7, and 9 × 9 were named SPP1, and the maximum pooling filter of 3 × 3, 5 × 5, 7 × 7, and 9 × 9 was named SPP2. These structures are shown in Fig. 7.

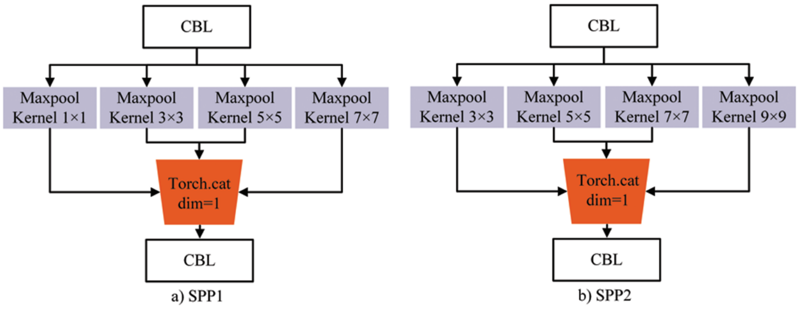


Figure 7: Improved spatial pyramid pooling micro structures: (a) SPP1; (b) SPP2

CBL is a combined module in Fig. 7 that consists of a convolution layer, a BN layer, and an activation function layer. The fusion of multiple receptive fields can be improved by adding a smaller maximum pooling layer, and the algorithm pays more attention to local information with a smaller maximum pooling layer, thus improving the detection accuracy of small targets. This paper uses the SPP1 module SPP module replacement of the original backbone network, at the same time respectively in the neck on the network for the first time after sampling increase SPP module, after the second sampling on SPP2 module, this according to the different characteristics of different output detection layer using pyramid pooling module, can enhance the characteristics of the corresponding output detection layer expression ability, achieve better detection effect.

3.3 Loss Function

The YOLOv5 algorithm's initial loss function is shown in formula (2). Two types of cross-entropy loss functions are used for confidence and class loss functions. For the position loss function, generalized intersection over union (GIOU) is utilized, and GIOU_Loss is shown in the formula (3).

$$\text{Loss} = \text{GIOU\_Loss} + \text{Loss}_{\text{conf}} + \text{Loss}_{\text{class}} \tag{2}$$

$$\text{GIOU\_Loss} = 1 - \text{IOU} + \frac{|Q|}{C} \tag{3}$$

where C represents the smallest boundary rectangle between the detected frame and the previous frame, and Q represents the difference between the smallest boundary rectangle and the addition of these two boxes. However, like intersection over union (IOU), GIOU only considers the overlap degree of two frames. The overlap part cannot be optimized and has certain limitations, as shown in Fig. 8 when the detection frame and the real frame contain each other.

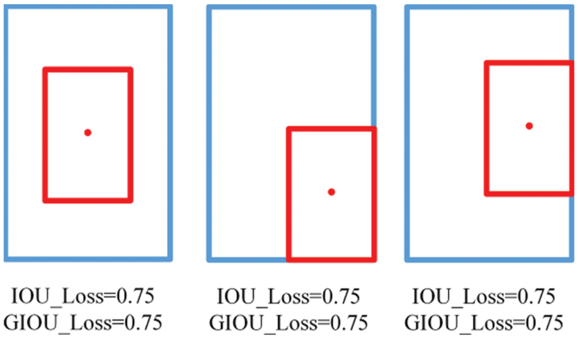


Figure 8: These situations in which a GIOU loss degrades to an IOU loss

Because of the above problems, we designed the RIOU_Loss function. RIOU_Loss function considers three aspects of the overlap area, center point distance, and the diagonal length, and is specifically defined as for formulas (4)–(6) below.

$$\text{RIOU\_Loss} = 1 - \text{IOU} + \frac{\rho^2 (p, g)}{C^2} + \frac{|R - r|}{C + |R - r|} \tag{4}$$

$$R = \sqrt{(w^g)^2 + (h^g)^2} \tag{5}$$

$$r = \sqrt{(w^p)^2 + (h^p)^2} \tag{6}$$

As shown in Fig. 9, $\rho(p, g)$ represents the Euclidean distance between the central points of the prediction frame and the labeling frame, p and g are the central points of the two frames, C represents the diagonal lengths of the minimum enclosing matrix frames of the two frames, the w^g and h^g are the width and height of the prediction frame respectively, and the w^p and h^p are the width and height of the true frame respectively, R is the diagonal lengths of the target frame, and r is the diagonal lengths of the prediction frame.

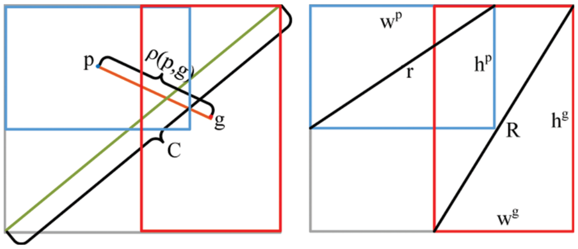


Figure 9: RIOU_Loss for bounding box regression

This limiting mechanism of increasing center point distance and diagonal length effectively avoids the problem that GIOU_Loss will produce a larger outer frame and loss value when the two frames are far apart, making the function convergence faster. At the same time, the c value is unchanged when the two boxes are in the inclusion relationship. However, the ρ(p, g) and R values will change, making the prediction box more consistent with the real box. Algorithm 1 shows the detailed steps of RIOU_Loss in our method:

Algorithm 1: Detailed process of RIOU_Loss function

```
input: Two arbitrary convex shapes: A, B
output: RIOU_Loss
Step1: For A and B, finds the smallest enclosing convex object c, finds the center points p and g
Step2: For c, find the diagonal length C
Step3: Find the diagonal lengths R and r according to formulas (5) and (6), respectively
Step4: Obtain RIOU_LOSS according to formula (4)
```

4 Evaluation Metrics

4.1 Experimental Datasets

To test the generalization ability of the model, two small aerial datasets, VEDAI and RSOD, are selected for testing in this experiment. In the experiment, the improved algorithm RT-YOLO was tested by the quantitative comparison method.

4.1.1 VEDAI Dataset

The public dataset of aerial image VEDAI proposed in 2014 is adopted, which contains 1210 RGB images with a resolution of 1024 × 1024 (or 512 × 512) pixels. The whole dataset includes 3640 instances including vehicles, ships and aircraft, including 9 categories of “car”, “truck”, “camping car”, “tractor”, “aircraft”, “ship”, “pick up”, “Van” and “others”. The number of each target is shown in Table 2, because all targets are distributed in fields and grasslands in areas with a rich background such as mountains and urban areas. There are an average of 5.5 vehicles per image, the image illumination has a great impact, and the direction of vehicle targets is random. As a public dataset for small object detection, VEDAI is challenging.

Table 2: Number of each target in the VEDAI dataset

Class	Boat	Camping	Car	Pickup	Tractors	Trunk	Vans	Airplane	Others
Number	170	390	1340	950	190	300	100	47	200

4.1.2 RSOD Dataset

The RSOD dataset was released by Wuhan University in 2015 and is mainly used in the field of object detection. It contains 976 remote-sensing images. The detected targets have different scales, orientations, and current situations. The dataset is marked with 6950 target location information, and the target categories are 4 categories, including aircraft, oil tank, overpass, and playground. Before the experiment, we round off 40 pictures in the playground category in the RSOD dataset. Finally, the total dataset was 936 pictures, which were divided into the training set and test set according to the ratio of 2:1:1, including 468, 234, and 234 small target images respectively. Table 3 lists the number of targets in each category contained in the training set and test set respectively.

Table 3: Number of targets in the RSOD dataset

Class	Train set	Validation set	Test set	total
Aircraft	2145	1248	1100	4993
Oil tank	834	362	390	1586
Overpass	88	46	46	180
Playground	97	47	47	191

4.2 Experiment Environment

We carry out experiments on two aviation small target datasets, VEDAI and RSOD compared with other most advanced object detection algorithms. The following are the experimental conditions: Python 38, Pytorch 1.7.0, GPU 11.0 Framework. Ubuntu is the operating system. CPU:i7-7700k. NVIDIA GeForce RTX 3080 graphics card. We select the coco datasets commonly used in the object detection task for pre-training, and set the parameter initialization for the model training: the size of the input picture is 640 × 640, the initial learning rate is 0.001 and the batch size is set to 32. The optimizer is the Adam algorithm, and the training epochs are 300 times. The initialization parameters are displayed in Table 4.

Table 4: The initialization parameters of training

Input size	Batch size	Momentum	Learning rate	Epoch
640 × 640	32	0.9	0.001–0.00001	300

4.3 Evaluation Index

There are two types of acknowledged performance evaluation indicators for object detection algorithms: evaluating algorithm detection accuracy and assessing algorithm detection speed. Precision (P), recall (R), average precision (AP), mean average precision (mAP), and other metrics are used to evaluate the algorithm's detection ability. The algorithm's detecting speed is primarily measured in frames per second (FPS). These public indicators are also used in this paper's Evaluation Metrics.

In a detection algorithm that can identify C-type objects, the images containing the i-th object are detected in $\mathrm{M_i}$ frames during the execution of the detection task, where $i \in \{1, 2, \ldots, C\}$, select the one with the highest confidence. The first N frames of images ($N \in \{1, 2, \ldots, M\}$), calculate the intersection and union ratio of the Bounding Box predicted by the algorithm in each image and the actual area corresponding to the target. The images whose intersection and the union ratio are greater than a certain threshold are classified as true positive (TP) with accurate prediction. The number is represented by TP_i^N, and the number of false positive (FP) images with the wrong prediction is represented by FP_i^N, TN_i^N represent the number of true negative (TN) prediction mistakes. Precision (P) is defined as the ratio of True Positives in the first N photos used to identify the target (i), and the formula is as follows:

$$P_i = \frac{\mathrm{TP}_i^N}{N} = \frac{\mathrm{TP}_i^N}{\mathrm{TP}_i^N + \mathrm{FP}_i^N} \tag{7}$$

The ratio of True Positives to the total number of image frames that contains the i-th object is known as recall (R). If it is assumed that K image frames containing objects of the i-th type exist. The following is the formula for calculating it:

$$R_i = \frac{\mathrm{TP}_i^{\mathrm{N}}}{K} = \frac{\mathrm{TP}_i^N}{\mathrm{TP}_i^N + \mathrm{TN}_i^N} \tag{8}$$

Average accuracy (AP) is also often used to quantitatively assess the performance of detection algorithms. Intuitively, AP is the area under the P-R curve. Generally speaking, the better the classifier, the higher the AP value. Further, the most important indicator in the object detection algorithm, mAP, can be obtained by averaging the AP of each category. The size of mAP must be in the [0, 1] interval, and the higher the index, the better the global accuracy of the algorithm. The formula for calculating mAP is as follows:

$$\mathrm{AP} = \int_0^1 P(R)\, dR \tag{9}$$

$$\mathrm{mAP} = \sum\nolimits_{i=1}^{C} \mathrm{AP}_i / C \tag{10}$$

where AP is the average accuracy of a single target category, and C represents the number of classifications.

The detection speed is a significant indicator of the measuring algorithm's evaluation. It's critical to pay particular attention to whether the algorithm can analyze enough video frames/images promptly, for time-sensitive needs for high real-time detection systems. The number of frames per second (FPS) is the rate at which a detection system completes a object detection task in a given amount of time (s). FPS is a standard metric for determining the algorithm's detection speed.

5 Results and Discussion

We analyze the backbone network replacement, method ablation and different model comparison to reflect the effectiveness of the improved methods.

5.1 Experiment Comparison of the Improved Backbone

In the original YOLOv5 algorithm, the detection of the UAV aerial image dataset is often missed and misdetected. This may be caused by the poor ability of the backbone network to extract the features of the small targets, and the detection scale fails to match the scale size of the small targets in the image. On the one hand, the original CSP structure uses hierarchical multi-scale to utilize features, and does not fully utilize the internal features of a single layer. We design Res2T to enhance fine-grained monolayer Internal feature utilization. On the other hand, the original CSP structure consistently valued the characteristics of each channel. We increase triplet attention modules to establish relationships between dimensions in enhanced spatial and channel information quality. Considering the computational load of the algorithm. We remove the convolutional layers on the short branches of the original module. To verify the effectiveness of the modified backbone network, we perform the ablation experiments on the VEDAI datasets using Darknet53, CSPDarknet53, ResNet50, ResNet101, VGG16, and RepVGG. The experimental results are given in Table 5.

Table 5: Comparison experiment of the backbone network

Model	Backbone	Input size	mAP (%)	FPS
YOLOv5	Darknet53	640 × 640	48.6	29.5
YOLOv5	CSPDarknet53	640 × 640	57.2	31.2
YOLOv5	ResNet50	640 × 640	41.5	31.1
YOLOv5	ResNet101	640 × 640	43.4	19.8
YOLOv5	VGG16	640 × 640	50.8	16.8
YOLOv5	RepVGG	640 × 640	54.5	49.2
YOLOv5	Ours	640 × 640	59.1	30.3

On the same input dimension, the mAP of our method on the test data set exceeds the original csparknet53 network, 9.5% higher than darknet53%, and 17.6% and 15.7% higher than the other two residual networks resnet50 and resnet101, respectively. Compared with the newly proposed RepVGG, our method also has a better detection effect. The experimental results show that the improved backbone network. Therefore, strengthening the connection between internal feature utilization and enhanced channels can effectively improve the extraction ability of small target features, improve the detection performance of the algorithm for small targets, and effectively reduce the error detection of small targets in UAV aerial image tasks.

5.2 Ablation Experiment

To verify the impact of the improved method on the detection power of the YOLOv5 algorithm, we perform the ablation experiments on the VEDAI dataset, and the results are shown in Table 6. Where B represents the improved backbone network, N represents changing the SPP pooling in the neck network, and L represents the use of the improved loss function.

Table 6: Ablation experiment of the improved module

Model	B	N	L	Input size	mAP (%)	FPS
YOLOv5				640 × 640	57.2	31.2
YOLOv5	√			640 × 640	59.1	29.7
YOLOv5	√	√		640 × 640	60.5	29.4
YOLOv5	√	√	√	640 × 640	60.8	30.3

After the addition of the improved Backbone, the mAP value of the model is increased from 57.2% to 59.1%. This demonstrates that improved backbone networks can effectively enhance the feature utilization power of the model for small targets. In addition, the unusual size SPP structure is used in the neck network to improve the binding ability of the receptive field, which increases the mAP of the network detection test set from 59.1% to 60.5%. Finally, the improved RIOU loss function is used to accelerate the network convergence and improve the network detection accuracy by 0.3%. Experimental results show that the improved RT-YOLO algorithm can effectively improve the tiny target detection of UAV aerial images.

5.3 Comparative Experiment of Different Models

To validate the improved method, we compare it with other state-of-the-art algorithms on the VEDAI dataset and the RSOD datasets.

5.3.1 VEDAI Dataset Comparison Experiments

In the experiment of the VEDAI dataset, the training set and testing set are divided into 4:1. In the training phase, 994 images in the dataset are selected as training samples, and in the test phase, the remaining 248 images are selected as test samples. Comparing several mainstream object detection frameworks, the specific experimental results are shown in Table 7.

Table 7: The VEDAI dataset experimental results

Model	AP (%)										FPS
	Boat	Camping	Car	Pickup	Tractor	Trunk	Van	Airplane	Other	mAP	
Faster RCNN	56.6	75.4	73.9	69.7	73.4	46.4	83.9	77	43.1	65.7	6.3
SSD	31.6	51.3	67.6	50.4	43.5	45.1	42.7	61.5	22.3	46.1	32.0
YOLOV3	64.5	54.7	69.1	52.8	52.5	32.8	51.3	51.6	31.4	51.3	29.5
YOLOV4	77.1	71.9	86.7	75.2	72.7	68.1	52.0	81.0	43.0	72.5	25.7
YOLOv5s	76.9	57.9	72.1	56.8	48.2	42.5	51.3	51.2	33.8	57.2	31.2
RT-YOLO	76.6	54.2	75.2	61.2	52.1	48.2	60.6	63.5	44.1	60.8	30.3

From the analysis of Table 7, it can be seen that the mean mAP of the dataset increased from 57.2% to 60.8%, which is due to the improvement of the RT-YOLO algorithm using a better backbone network and higher resolution images as input. In Table 7, the object detection performance of Tractors, Van, Trunk, Airplanes, and Other is improved greatly, and the average accuracy of Airplanes was increased from 51.2% to 63.5%, with an increase of 12.3%. This shows that the improved algorithm for aerial image target detection performance has been greatly improved. Boat, Camping detection accuracy is slightly reduced, is due to the RT-YOLO algorithm using Triple Attention in reducing compute operations at the same time, to a certain extent weakening the expressive ability of convolution, thus to some extent affecting the part of the single layer feature change insensitive category target detection accuracy. From the comparison of experimental results, we can see that the improved algorithm has achieved a better detection effect. Fig. 10 shows the comparison of the test results of the VEDAI dataset. The comparison shows that RT-YOLO can detect smaller objects on the VEDAI dataset more efficiently.

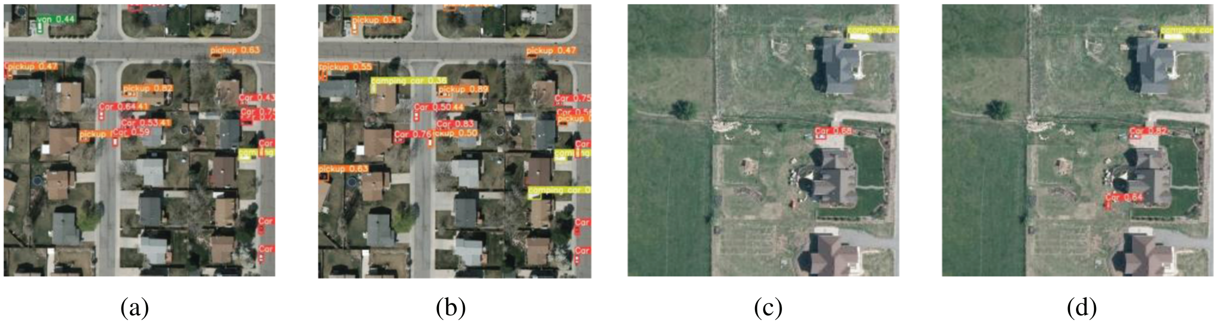


Figure 10: Detection results of the VEDAI dataset: (a) and (c) is the detection results of YOLOv5; (b) and (d) is the detection results of RT-YOLO

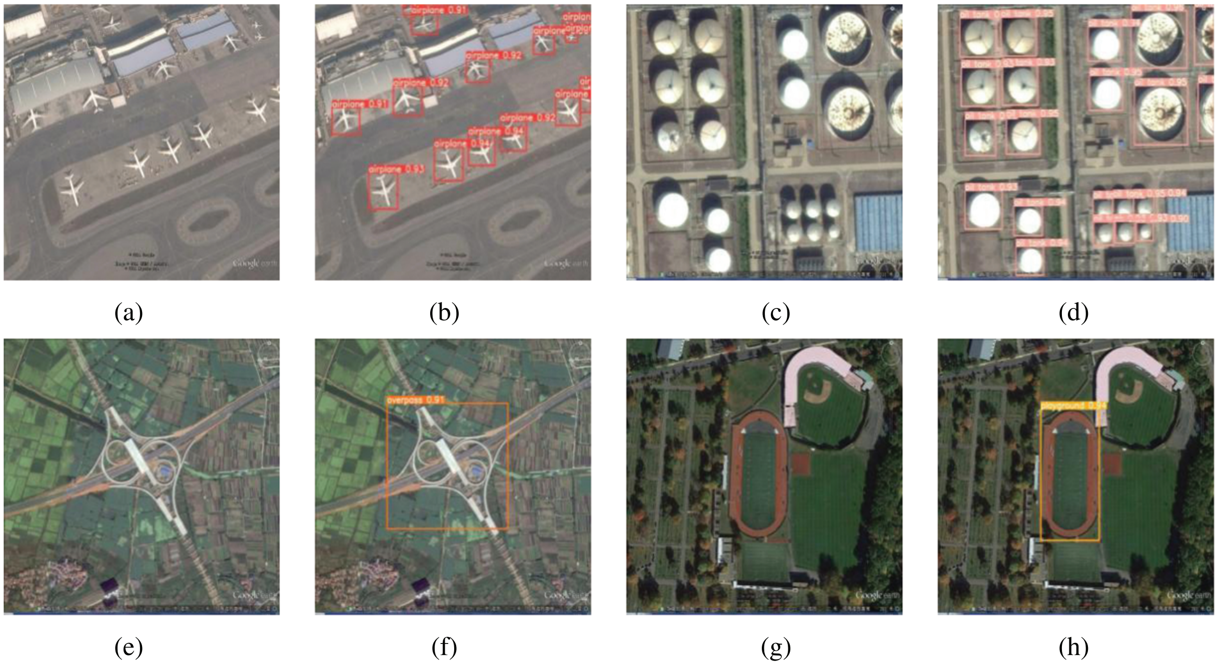
5.3.2 RSOD Dataset Comparison Experiments

Table 8 lists the object detection accuracy of the improved algorithm in this paper on the RSOD dataset. From the total average accuracy of various types, the method in this paper is higher than other algorithms except for yolov4. From the detection accuracy of the single category algorithm, our method is higher than the original yolov5s algorithm in aircraft, oil tanks, and overpasses. Through comparison, it can be seen that our method performs better in small object detection.

Table 8: The comparative results of different categories in the RSOD dataset

Model	AP (%)					FPS
	Airplane	Oil tank	Overpass	Playground	mAP	
Faster RCNN	83.5	98.1	88.6	97.8	92.0	9.6
SSD	71.8	90.7	90.2	98.5	87.8	42.2
YOLOV3	89.7	96.5	80.9	96.8	91.6	29.7
YOLOV4	92.3	98.9	86.9	99.5	95.2	28.2
YOLOV5	93.6	98.5	83.8	98.7	93.7	52.6
RT-YOLO	94.5	99.4	86.3	97.1	95.4	49.4

Fig. 11 shows the detection results of our method on the RSOD dataset. Through the corresponding comparison between the original graph and the detection result graph, it can be concluded that the detection performance of our method is relatively excellent.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 11: Detection results of RSOD dataset: (a), (c), (e), and (g) is the original image of the RSOD dataset; (b), (d), (f), and (h) represent the detection results of the RT-YOLO

6 Conclusion and Prospect

To improve the detection accuracy of aerial small object detection tasks, a new small object detection algorithm for aerial image is proposed in this paper. The algorithm is called RT-YOLO by us. We design a new feature extraction network structure CSP_RT, integrates the triplet attention mechanism to improve backbone network. In order to improve the sensitivity of RT-YOLO to small target, we design a new space Pyramid pooling SPP1 module and SPP2 module optimize the receptive field fusion. Considering the overlap area, center point distance, and aspect ratio, we propose the RIOU_Loss loss function. Using the VEDAI dataset, research, analysis, and proof of the attention mechanism to improve the performance of the small object detection algorithm, and found that adding the SPP model to the network neck is more friendly to the extraction of small target feature information. Experiments have proved that RT-YOLO effectively improves the detection accuracy of small object aerial image. The mAP@0.5 value on the VEDAI test set is increased by 3.6% compared with YOLOv5s, and the mAP@0.5 on the RSOD dataset is increased by 1.7%.

However, our method of integrating the attention module and adding and improving the SPP module will increase the number of algorithm parameters and floating-point operations, and reduce the real-time performance of algorithm detection. In the next research, we will consider compressing and pruning the model to lighten the network model and improve the real-time performance of model detection based on ensuring the accuracy of algorithm detection.

Acknowledgement: The author would like to thank the support of the Scientific Research Project of the Hunan Provincial Department of Education and the Science and Technology Plan Project of Hunan Province.

Funding Statement: This work was supported in part by the Scientific Research Project of Hunan Provincial Department of Education under Grant 18A332 and 19A066, authors HW.D and Z.C, http://kxjsc.gov.hnedu.cn/; in part by the Science and Technology Plan Project of Hunan Province under Grant 2016TP1020, author HW.D, http://kjt.hunan.gov.cn/.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Hu and G. H. Lee, "Image-based geo-localization using satellite imagery," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1205–1219, 2020.

[2] K. Gao, B. Liu, X. Yu, P. Zhang, X. Tan *et al.,* "Small sample classification of hyperspectral image using model-agnostic meta-learning algorithm and convolutional neural network," *International Journal of Remote Sensing*, vol. 42, no. 8, pp. 3090–3122, 2021.

[3] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, pp. 580–587, 2014.

[4] R. Girshick, "Fast R-CNN," in *2015 IEEE Int. Conf. on Computer Vision (ICCV)*, Santiago, Chile, pp. 1440–1448, 2015.

[5] S. Ren, K. He, R. Girshick and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[6] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask r-cnn," in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 2961–2969, 2017.

[7] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779–788, 2016.

[8] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 7263–7271, 2017.

[9] S. Zhang, L. Chai and L. Jin, "Vehicle detection in UAV aerial images based on improved YOLOv3," in *2020 IEEE Int. Conf. on Networking, Sensing and Control (ICNSC)*, Nanjing, China, pp. 1–6, 2020.

[10] J. H. Sejr, P. Schneiderkamp and N. Ayoub, "Surrogate object detection explainer (SODEx) with YOLOv4 and LIME," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 662–671, 2021.

[11] Z. Wang, Y. Zheng, X. Li, X. Jiang, Z. Yuan *et al.,* "DP-YOLOv5: Computer vision-based risk behavior detection in power grids," in *2021 Chinese Conf. on Pattern Recognition and Computer Vision (PRCV)*, Zhuhai, Guangdong, China, pp. 318–328, 2021.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "Ssd: Single shot multibox detector," in *2016 European Conf. on Computer Vision (ECCV)*, Amsterdam, Netherlands, pp. 21–37, 2016.

[13] Y. Zhang, W. Zhou, Y. J. Wang and L. J. Xu, "A real-time recognition method of static gesture based on DSSD," *Multimedia Tools and Applications*, vol. 79, no. 4, pp. 17445–17461, 2020.

[14] S. Bera, V. K. Shrivastava and S. C. Satapathy, "Advances in hyperspectral image classification based on convolutional neural networks: A review," *Computer Modeling in Engineering & Sciences*, vol. 133, no. 2, pp. 219–250, 2022.

[15] L. W. Sommer, T. Schuchert and J. Beyerer, "Fast deep vehicle detection in aerial images," in *2017 IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, pp. 311–319, 2017.

[16] X. Yu, Y. Gong, N. Jiang, Q. Ye and Z. Han, "Scale match for tiny person detection," in *2020 IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, pp. 1246–1254, 2020.

[17] X. Yang, J. Yang, J. Yan, T. Zhang, Z. Guo *et al.,* "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *2019 IEEE Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 8232–8241, 2019.

[18] A. Ibrahim, S. Mirjalili, M. El-Said, S. S. M. Ghoneim, M. Al-Harthi *et al.,* "Wind speed ensemble forecasting based on deep learning using adaptive dynamic optimization algorithm," *IEEE Access*, vol. 9, pp. 125787–125804, 2021.

[19] Y. Rao, H. Mu, Z. Yang, W. Zheng, F. Wang *et al.,* "B-PesNet: Smoothly propagating semantics for robust and reliable multi-scale object detection for secure systems," *Computer Modeling in Engineering & Sciences*, vol. 132, no. 3, pp. 1039–1054, 2022.

[20] Y. Yang, N. Zhu, Y. Wu, J. Cao, D. Zhan *et al.,* "A semi-supervised attention model for identifying authentic sneakers," *Big Data Mining and Analytics*, vol. 3, no. 1, pp. 29–40, 2020.

[21] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *2018 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 3578–3587, 2018.

[22] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang *et al.,* "Lite-hrnet: A lightweight high-resolution network," in *2021 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 10440–10450, 2021.

[23] X. Xu, T. Gao, Y. Wang and X. Xuan, "Event temporal relation extraction with attention mechanism and graph neural network," *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 79–90, 2022.

[24] W. Chen, M. Liu, X. Zhou, J. Pan and H. Tan, "Safety helmet wearing detection in aerial images using improved YOLOv4," *Computers Materials & Continua*, vol. 72, no. 2, pp. 3159–3174, 2022.

[25] Y. Li, F. Liu, S. Wang and J. Yin, "Multi-scale symbolic lempel-ziv: An effective feature extraction approach for fault diagnosis of railway vehicle systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 1, pp. 199–208, 2020.

[26] J. Su, D. Yi, B. Su, Z. Mi and W. H. Chen, "Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2242–2249, 2020.

[27] W. Zhao, Z. Zhang, S. Mirjalili, L. Wang, N. Khodadadi *et al.,* "An effective multi-objective artificial hummingbird algorithm with dynamic elimination-based crowding distance for solving engineering design problems," *Computer Methods in Applied Mechanics and Engineering*, vol. 398, no. 15, pp. 115–223, 2022.

[28] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng *et al.,* "Perceptual generative adversarial networks for small object detection," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 1951–1959, 2017.

[29] Y. Bai, Y. Zhang, M. Ding and B. Ghanem, "Sod-mtgan: Small object detection via multi-task generative adversarial network," in *2018 Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 206–221, 2018.

[30] D. K. Das, S. Shit, D. N. Ray and S. Majumder, "CGAN: Closure-guided attention network for salient object detection," *The Visual Computer*, vol. 38, no. 11, pp. 3803–3817, 2022.

[31] T. Zhao, X. Yi, Z. Zeng and T. Feng, "MobileNet-yolo based wildlife detection model: A case study in yunnan tongbiguan nature reserve, China," *Journal of Intelligent & Fuzzy Systems*, vol. 41, no. 1, pp. 2171–2181, 2021.

[32] H. Pan, D. Badawi and A. E. Cetin, "Fourier domain pruning of MobileNet-v2 with application to video based wildfire detection," in *2020 25th Int. Conf. on Pattern Recognition (ICPR)*, Milan, Italy, pp. 1015–1022, 2020.

[33] H. Wang, V. Bhaskara, A. Levinshtein, S. Tsogkas and A. Jepson, "Efficient super-resolution using mobilenetv3," in *2020 European Conf. on Computer Vision (ECCV)*, Glasgow, US, pp. 87–102, 2020.

[34] X. Zhang, X. Zhou, M. Lin and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *2018 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 6848–6856, 2018.

[35] J. Dong, J. Yuan, L. Li, X. Zhong and W. Liu, "An efficient semantic segmentation method using pyramid ShuffleNet V2 with vortex pooling," in *2019 31st IEEE Int. Conf. on Tools with Artificial Intelligence (ICTAI)*, Portland, OR, USA, pp. 1214–1220, 2019.

[36] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *Journal of Visual Communication and Image Representation*, vol. 34, no. 1, pp. 187–203, 2016.

[37] Y. Long, Y. Gong, Z. Xiao and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2486–2498, 2017.

[38] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye *et al.,* "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 8574–8586, 2022.

[39] S. Gao, M. M. Cheng, K. Zhao, X. Zhang, M. Yang *et al.,* "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.