Hindawi
Advances in Multimedia
Volume 2022, Article ID 7234888, 10 pages
https://doi.org/10.1155/2022/7234888

Research Article

Study on the Lightweighting Strategy of Target Detection Model with Deep Learning

Junli Hu

Henan Industry and Trade Vocational College, Zhengzhou 451191, China

Correspondence should be addressed to Junli Hu; hujunli@hngm.edu.cn

Received 26 May 2022; Revised 2 July 2022; Accepted 18 July 2022; Published 30 August 2022

Academic Editor: Qiangyi Li

Aiming at the high miss detection and false detection rate of traditional SSD (single shot multibox detector) target detection algorithm in target detection, this paper proposes a lightweight detection algorithm for deep learning target detection model in order to improve the detection accuracy. Firstly, the real-time and efficient target detection backbone network VoVNet is used to replace the feature extraction network VGG16. The residual structure is integrated to solve the problem of VoVNet network degradation to improve network performance. Secondly, self-attention mechanism is introduced to capture multiscale local and global information to obtain richer image semantic features. According to the characteristics of the target sample size, the anchor frame is designed by using a priori information. In the network training, the anchor frame by selection enhancement based on IoU optimization is used to fully train the target information and strengthen the reading ability of the network to small-scale targets. Experiments on the public data set AI-TOD to show that the target detection lightweight model of deep learning has stronger detection ability and higher average detection accuracy than other algorithms, which proves the applicability and effectiveness of this algorithm.

1. Introduction

Target detection is to identify the target of interest in the image and determine the location, which is widely used in various fields of life. In practical applications, a large number of targets are defined as small targets due to their size or distance. The pixel coverage is small and the ability to express features is not enough. In the current depth conversion accuracy model, features are easily lost, which makes small target detection difficult to conventional target detection [1].

Researchers have done a lot of exploration in solving small target detection, mainly including anchor box optimization, introducing attention mechanism, feature fusion, feature enhancement, improving network, improving loss function, and so on [2]. By adding a bottom-up path to FPN, literature [3] shortens the upward transmission distance of location information and realizes the effective transmission of location information to a deep feature map. Because the small target occupies too few pixels and the information that can be directly used is not rich, researchers strengthen the ability of model detection through context information. Literature [4] adds a deconvolution layer to obtain context information based on SSD model and generates a "wide narrow wide" hourglass structure. Literature [5] introduced a recursive reverse path from deep to shallow into SSD model and enhanced the semantic features containing context information in the deep layer to the shallow layer through a feature enhancement module. Literature [6] designed a Trident Net using hole convolution to use the semantic information of the target context. Literature [7] proposes a lightweight and efficient feature fusion module based on SSD, which makes full use of features, greatly improves the accuracy under the condition of less loss of detection speed, and has a better effect on small target detection. Starting from the feature pyramid network (FPN), literature [8] adds a fusion factor to the FPN to describe the coupling degree of adjacent layers to control the deep transmission of shallow information, which makes the FPN more suitable for small targets and improves the performance of small target detection. Literature [9] proposed a

copy enhancement method to solve the problems of small target coverage, lack of diversity of occurrence positions, and the intersection and union ratio between detection frame and truth box is far less than the expected threshold. By copying and pasting small targets in the image many times, the number of training samples of small targets is increased, so as to improve the detection performance of small targets. Reference [10] proposed an adaptive resampling strategy for data enhancement in RRNet. This strategy copies the target image considering the context information based on the pretrained semantic segmentation network, so as to solve the problems of background mismatch and scale mismatch that may occur in the process of simple replication, so as to achieve a better data enhancement effect. Literature [11] starts from the problems of small proportion of the number of targets and less information contained in itself and zooms and stitches the images in the training process. Reference [12] proposes a multilevel feature fusion algorithm, which balances the speed and accuracy of small target detection. Reference [13] proposes a single time multi box detector for feature fusion, which uses a lightweight feature fusion module to connect and fuse the features of each layer to a larger scale, and then constructs a feature pyramid on the obtained feature map for detection, which improves the detection performance of small targets at the expense of less speed. Reference [14] proposed a small target detection method of airport pavement combining multiscale feature fusion and online difficult case mining. This method uses ResNet-101 as the feature extraction network and establishes a "top-down" feature fusion module with up sampling based on the network to generate a high-resolution feature map with richer semantic information.

Aiming at the shortcomings of traditional algorithms and combined with the characteristics of SSD algorithm, this paper proposes a deep learning target detection lightweight model. Experimental results show that the fusion algorithm improves the target detection accuracy and verifies the effectiveness of the improved algorithm. The average detection accuracy on the target data set AI-TOD is higher than that of other mainstream feature extraction networks, which better solves the problem of difficult target detection. It can be applied to all kinds of target detection models to replace the original feature extraction network and adapt to the task of target detection.

The innovations and contributions of this paper are as follows:

(1) The real-time and efficient target detection backbone network VoVNet is used to replace the feature extraction network vgg16, which reduces the parameters of the network model and reflects the network lightweight.

(2) According to the size characteristics of the target sample, the anchor frame is designed by using a priori information.

Self-attention mechanism is introduced to capture multiscale local and global information to obtain richer image semantic features.

This paper consists of five main parts: the first part is the introduction, the second part is the state of the art, the third part is the system design of this paper, the fourth part is the experiment and analysis, and the fifth part is the conclusion, besides there are abstracts and references.

2. State of the Art

2.1. Definition of Small Target. Different scenarios have different definitions of small targets. The existing definition methods of small targets are mainly divided into the following two categories:

(1) Based on relative scale definition. That is, the small target is defined from the perspective of the relative proportion between the target and the image. The ratio between the width and height of the target bounding box and the width and height of the image is less than a certain value, and the more general ratio value is 0.1. The ratio of target bounding box area to image area is less than a specific value, and the broader value is 0.03. The small target is defined according to the proportion between the actual coverage pixel of the target and the total pixel of the image. However, these definitions based on the relative scale have many problems. For example, this definition method is difficult to effectively evaluate the detection performance of the model for targets with different scales. In addition, this definition method is easily affected by data preprocessing and model structure.

(2) Based on absolute scale definition. The small target is defined from the perspective of the absolute pixel size of the target. At present, the most common definition comes from the general data set in the field of target detection, which defines a small target as a resolution of less than 32 pixels × 32 pixel target. Why 32 pixels × 32 pixels, this paper thinks from two directions. One idea comes from the research of literature [15]. Human beings can effectively recognize the scene on the image, and the required color image pixel size is 32 pixels × 32 pixels, i.e., less than 32 pixels × 32 pixel targets are difficult for humans to recognize. Another idea comes from the structure of convolutional neural networks in deep learning. Take VGG-Net, the classical network structure published in the same year as the first part of the MS COCO data set, as an example, which leads to the "point" on the final feature vector corresponding to the pixel size on the input image of 32 pixels × 32 pixels. Therefore, considering the difficulty of feature extraction, 32 pixels can be used × 32 pixels is used as a defining standard to distinguish small targets from conventional targets. In addition to MS COCO, other definitions are based on an absolute scale. For example, targets with pixel values between [10, 50] are defined as small targets in both aerial image data set DOTA and face detection data set WIDE FACE. In the pedestrian detection dataset CityPersons, the

small target is defined as the target with a height of less than 75 pixels for the pedestrian, which has a special proportion. Tiny Person, a small pedestrian data set based on aerial images, defines small targets as targets with pixel values between [20, 32], and further defines targets with pixel values between [2, 16] as small targets.

2.2. Challenges of Small Target Detection. The mainstream definitions of small targets have been briefly described in the previous article. These definitions show that small targets account for a small proportion of pixels and have the basic characteristics of a small coverage area and less information. At present, the main reasons for the high difficulty of small target detection and the challenges it faces are as follows.

2.2.1. Few Available Features. Whether from the definition based on an absolute scale or relative scale, small targets have the problem of low resolution. Low resolution leads to less visual information and is vulnerable to environmental factors.

2.2.2. High Positioning Accuracy. Because small targets occupy a small area in the image, it is more difficult to locate their bounding box, which makes it difficult to detect small targets.

2.2.3. Small Target Accounts for Less. In the field of target detection, most of the existing data sets are for large/mesoscale targets, and less attention is paid to the special type of small targets. Although small targets account for 31.62% of MS COCO, each image contains too many instances and the distribution of small targets is uneven. At the same time, small targets are not easy to label. On the one hand, it comes from that small targets are not easy to be paid attention to by humans in the image, so it is difficult to label them completely. On the other hand, small targets are more sensitive to labeling errors. In addition, the existing small target data sets are often targeted at specific scenes, such as literature [16] for images in the airfield of view, literature [17] for faces, literature [18, 19] for pedestrians, literature [20] for traffic lights, and literature [21] for musical notes. The networks trained with these data sets are not suitable for general small target detection. In general, the large-scale general small target data set is still in a lack of state, and the existing algorithms do not have enough a priori information to learn, resulting in the insufficient performance of small target detection.

2.2.4. Sample Problem. In order to locate the position of the target in the image, most of the existing methods generate a series of anchor boxes at each position of the image in advance. How to solve the imbalance of small target and large/mesoscale target samples caused by anchor box mechanism is also a major challenge.

2.2.5. Small Target Aggregation Problem. Small targets are prone to aggregation. When small target aggregation occurs, the small targets adjacent to the aggregation area will aggregate into a point after multiple down sampling, which will be reflected in the deep feature map, resulting in the inability of the detection model to distinguish. When similar small targets appear densely, the predicted bounding box may also filter a large number of correctly predicted bounding boxes due to the non-maximum suppression operation of post-processing, resulting in missed detection. In addition, if the boundary box is too close, it will also make it difficult for the boundary box to regress and the model to converge.

2.2.6. Reasons for Network Structure. Due to the characteristics of small targets, the existing algorithms generally perform poorly. Although the detector design without anchor frame is a new development trend, the existing network is still the detector based on anchor frame, which is very unfriendly to small targets. In addition, in the training process of the existing network, the small target has less contribution to the loss function due to the small proportion of training samples, which further weakens the learning ability of the network for small targets.

3. Methodology

3.1. Lightweight SSD Target Detection Algorithm and Its Improvement. SSD algorithm uses a variety of data enhancement methods, which have higher detection accuracy than YOLO, faster detection speed than faster R-CNN, and better robustness.

However, it is not effective to detect targets with complex backgrounds and small sizes. Image targets are characterized by small size and dense quantity, which bring a great difficulty to classification and detection. Limited by the structure of the algorithm, it is unable to detect some small-scale targets, resulting in missed detection, false detection, and other problems. In this paper, VoVNet is used to replace the feature extraction network VGG-16 in the original algorithm, which reduces the parameters of the network model and reflects the lightweight of the network. At the same time, the residual structure is fused to solve the problem of network degradation, and the self-attention mechanism is fused into the algorithm structure to enhance the extraction ability of key information so that more key features are concerned by the algorithm. The anchor frame is designed by using a priori information and enhanced structure is shown in Figure 1 by supplementary selection based on IoU optimization, so as to solve the problem of mismatch and uneven distribution of anchor frame laying. The improved algorithm.

3.2. Replace the Skeleton Network. The high-efficiency backbone network VoVNet gathers the shallow features in series, which not only washes out the deep feature information that the early feature map can carry but also retains the information in the original form through series. It has better and more diversified feature representation and

feature mapping of more receiving domains than vgg16. VoVNet is a one-shot aggregation (OSA) network, which is composed of continuous convolution layers. At the same time, it aggregates the subsequent characteristic maps, which can effectively capture different acceptance domains, retain the cascade strength, and greatly improve the computing efficiency of GPU. The structure is shown in Figure 2. F represents convolution and $\otimes$ represents cascade.

Although VoVNet has the characteristics of high efficiency and diversity in feature description and representation, it still has great limitations in optimization. When OSA modules are stacked in VoVNet, network degradation will occur.

The residual structure can be used to solve the problem of network degradation. In the residual structure, the original feature is $H(i)$ and the residual feature is $H(i)$, then the learned residual feature is represented as follows:

$$F(i) = H(i). \tag{1}$$

In order to obtain more powerful performance, the residual structure will also learn new features. Batch normalization (BN) and Leaky-ReLU are added to the initial residual structure. BN normalized output feature layer can accelerate the convergence speed of the network and solve the problem of network training failure caused by gradient explosion. VoVNet has better feature extraction ability and higher accuracy than vgg16. Therefore, firstly, replace the initial skeleton network vgg16 in SSD algorithm with a deeper number network, but relatively more effective VoVNet, and integrate the improved residual structure with VoVNet to obtain a skeleton network with higher accuracy and stronger performance. As shown in Figure 3.

To verify the network's performance, VGG16, ResNet, VoVNet, and VoVNet (ours) are tested on the ImageNet2012 data set, and the results are shown in Table 1.

It can be seen from Table 1 that the accuracy of the improved VoVNet is much higher than that of the SSD initial backbone network, and the number of floating-point operations per second is much smaller than that of the initial network. Therefore, the overall performance of the network model in this paper is better than VGG16.

3.3. Introduce Self-Attention Mechanism. In essence, self-attention mechanism allocates different weights according to the importance of different information, so as to extract important feature information. In computer vision, insufficient semantic information will reduce the detection effect, but the self-attention mechanism can solve this defect. After capturing the global information, select the more critical feature information, and the stronger relevant features are convoluted and extracted, so as to better identify the detection target.

As shown in Figure 4, the feature map i is convolved through three 1×1 convolution kernels in two dimensions, b and m, before it is carried out to become $b \times m$, and is first convolved in two 1×1 convolution kernels in $b \times m$. The number of channels c_1 of the feature map is one-eighth of the initial number of channels, and the target feature space $f(i)s$ ($c_1 = c/8$, dimension $c_1 \times b \times m$) and the feature space $a(i)$ (dimension $c_1 \times b \times m$), matrix $f(i)$, and matrix $a(i)$ are used to extract pixel features and global features, respectively, for the third convolution with no reduction in the number of channels, and the matrix $b(i)$ of dimension $c \times b \times m$ is obtained. After the matrix multiplication of f and α, the scale feature matrix S_{xy} is obtained, and then the scale normalization is performed by the softmax function to obtain the attention matrix β_{xy}.

$$\begin{aligned} \beta_{xy} &= \frac{\exp\left(S_{xy}\right)}{\sum_{x=1}^{T} \exp\left(S_{xy}\right)}, \\ S_{xy} &= f(i_x)^N a(i_x), \end{aligned} \tag{2}$$

β_{xy} represents the degree of attention of the model to the position of each element in the attention matrix. S_{xy} represents the attention of the model to each element in the scale matrix, T represents the number of elements in the scale feature matrix, y represents the region, and x represents the position. The feature space b and the attention matrix are multiplied to obtain the self-attention feature map. The output attention layer is represented as follows:

$$\begin{aligned} O &= (O_1, O_2, O_3 \ldots O_I \ldots O_T) \in R^{C \times T}, \\ O_x &= \sum_{x=1}^{T} \beta_{y,x} b(i_x), \\ b(i_x) &= M_b i_x. \end{aligned} \tag{3}$$

In order to output the new feature map, add the weight to the self-attention feature map, and then input the self-attention feature map with the weight into the input feature map. The original weight value is 0, and the iteration of the weight value is completed by the back-propagation neural network, as shown below.

$$j_x = \gamma O_x + i_x, \tag{4}$$

j represents the final returned feature map, γ represents the initial value weight in the self-attention feature map, the initial value is 0, O_x represents the expanded self-attention feature map, and i_x represents the original input feature map.

3.4. Anchor Frame Design Based on Prior Information and by Selection Enhancement of IoU Optimization. The detection accuracy of SSD algorithm is directly affected by the scale proportion and size of anchor frame. If the size proportion is inappropriate, the detection accuracy will be directly reduced. The ideal range of anchor frame size is 20%~30% of the theoretical size. Increasing the gradient return of backpropagation and accelerating the rapid convergence of loss function simplifies the calculation of network forward propagation loss.

The window adjustment algorithm is used for type conversion of the data to fully extract the effective structure information, extract the prior information from the processed data, and design the anchor box to accelerate the

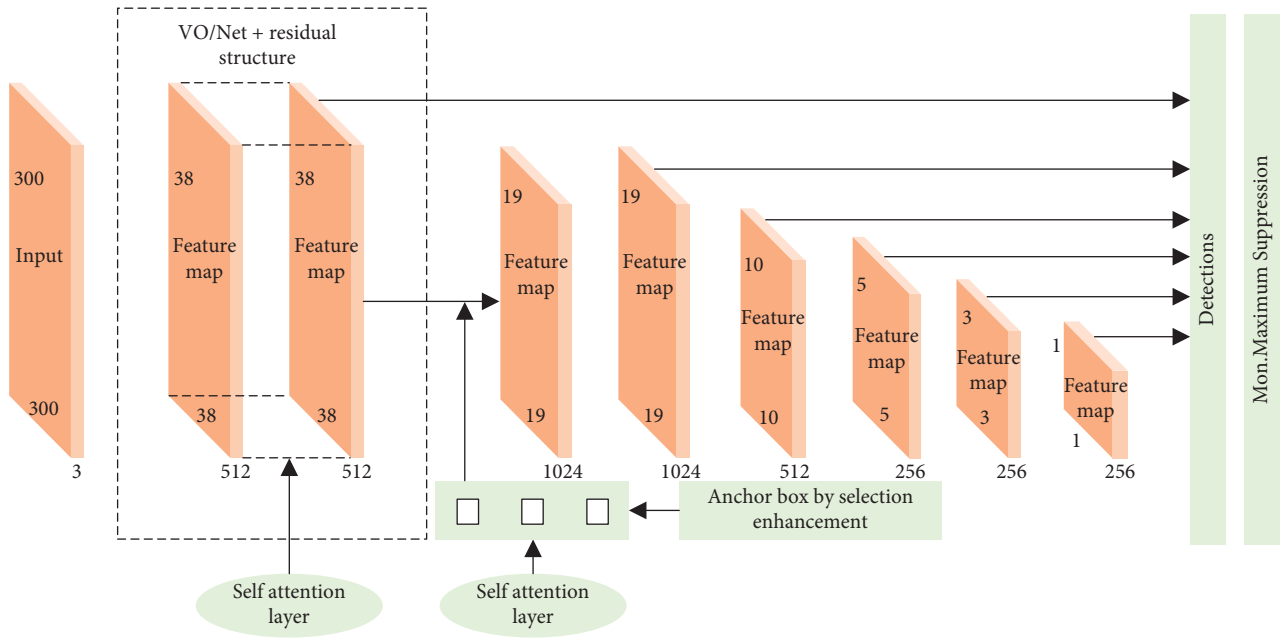


Figure 1: Improved SSD algorithm network structure diagram.

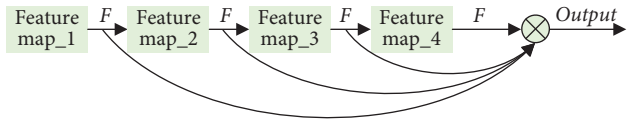


Figure 2: VoVNet structure diagram.

horizontal convergence of the prior information, so as to complete the accurate annotation finally. Extracting the prior information includes setting the initial curve and constructing the discriminant function, setting the initial curve according to the relationship between the data and the gray information of the image, speeding up the horizontal convergence speed, extracting the gradient features and texture features of the windowed image, constructing the discriminant function by machine learning, and finally using the prior information to achieve the most appropriate anchor frame annotation effect.

According to literature [22], the aspect ratio of image target is between 0.6–1.6. The target size recognized by feature map1 is very small. First set the aspect ratio of the anchor frame to 1. Randomly select the target in 30% of the image in the target data set, measure its target aspect ratio, and set the aspect ratio of the anchor frame of the feature map to $r \in \{1, 2/3, 3/2, 3/5, 5/3\}$ based on its prior information.

The density of anchor frame is calculated according to the anchor frame scale and down sampling multiple of each characteristic layer. d is the laying density of anchor frame, c is the dimension of anchor frame, and u is the lower sampling multiple. The laying density formula of anchor frame is as follows:

$$d = \frac{c}{u}. \tag{5}$$

After redesigning the anchor frame, the anchor frame selected in training the network directly affects the detection accuracy. When detecting image targets, the target size is inconsistent, the number of targets is large, and the regression is difficult, which is easy to cause the loss of targets. Therefore, after redesigning the anchor frame, IoU optimized anchor frame by selection enhancement is adopted to solve this kind of problem to the greatest extent.

As shown in Figure 5, after sorting the classification confidence, the convolutional neural network loses the anchor points with low classification confidence, but the anchor points with a similar confidence threshold will also be discarded. Therefore, there will be errors, which is easy to lead to wrong judgment of the network. The feature mapping points in the real box are mapped to the original input, a false box of the same size as the real box is set, and the unit between the two boxes is calculated. IoU is a pseudo box based on each point assigned, named IoU metric.

Let the real frame area be G, the false frame be H, S_* be the intersection area of G and H, and $S_G + S_H - S_*$ be the Union area of G and H.

Suppose l_H is the distance from the midpoint of the false frame to the left of the frame, r_H is the distance from the midpoint of the false frame to the right of the frame, n_H is the distance from the midpoint of the false frame to the top of the frame, h_H is the distance from the midpoint of the false frame to the bottom of the frame, similarly, l_G is the distance from the midpoint of the real frame to the left of the frame, r_G is the distance from the midpoint of the real frame to the right of the frame, n_G is the distance from the midpoint of the real frame to the top of the frame, and h_G is the distance from the midpoint of the real frame. The formula is as follows:

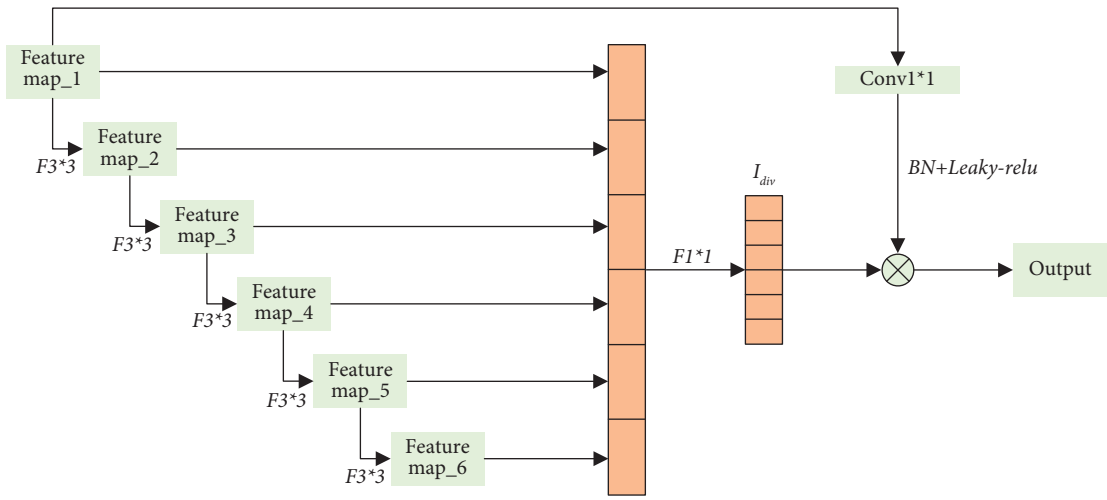


Figure 3: Improved VoVNet architecture diagram.

Table 1: Different network performance test results.

Network	Accuracy (%)	Flops (10^9)
VGG16	71.5	15.6
ResNet	73.2	3.4
VoVNet	74.1	1.9
VoVNet (ours)	74.7	1.7

$$\begin{aligned} l_G &= r_G \\ &= \frac{(l_H + r_H)}{2}, \\ n_G &= h_G \\ &= \frac{(n_H + h_H)}{2}, \\ S_H &= (l_H + r_H) \times (n_H + h_H), \\ S_G &= (l_G + r_G) \times (n_G + h_G), \\ S_* &= [\min(l_G, l_H) + \min(r_G, r_H)] \times [\min(n_G, n_H) \\ &\quad + \min(h_G, h_H)], \\ IoU &= \frac{S_*}{S_G + S_H - S_*}. \end{aligned} \tag{6}$$

In each network iteration, a part of the eliminated anchor box will continue to be added to the subsequent iterative training, which reduces the probability of small target loss and improves the accuracy of detection after many times.

4. Result Analysis and Discussion

4.1. Data Sets and Evaluation Indicators. Data set: in this paper, the data set AI-TOD used in aerial images is selected as the benchmark data set for model training and detection. It has 8 types such as vehicles and ships, including 28036 pictures, with a total of 700621 detection examples. The actual target size of AI-TOD data set is only 12.8 pixels on average, which is much smaller than other data sets, which is suitable for the research of this paper.

Evaluation index: in this paper, average precision (AP) is used as the evaluation index, including mAP, AP_{50}, AP_{75}, AP_s, and AP_m. AP_{50} represents the average precision value when the threshold of the real frame of the target and the predicted frame to union ratio (IoU) of the model is 0.5, AP_{75} is 0.75. mAP represents the equal distance between the threshold of the cross-to-union ratio from 0.5 to 0.95, take 10 values and calculates the average value of AP under these 10 thresholds. AP_s indicates that the pixels occupied by the detection target are less than 32^2 pixels, indicating that the pixels occupied by the detection target are between 32^2 and 96^2.

4.2. Experimental Environment and Parameter Setting. The configuration environment used in this experiment is shown in Table 2.

The experimental parameters are set as follows: the AI-TOD data set is used to train and test the model, and the input picture size is scaled to 416×416, use dual graphics cards for parallel training, and set the batch size of each graphics card to 8. The experiment adopts the random gradient descent algorithm, the initial value of the learning rate is set to 0.1, and a total of 350 epochs are trained. Before training, the image is flipped and trimmed, and other data enhancement operations are carried out. The same parameter settings are adopted for different models, and the experimental results are compared and analyzed.

4.3. Analysis of Experimental Results

4.3.1. Comparative Experiment of Different Network Structures. In order to verify the performance of the proposed network structure, the proposed feature extraction network structure is compared with the commonly used

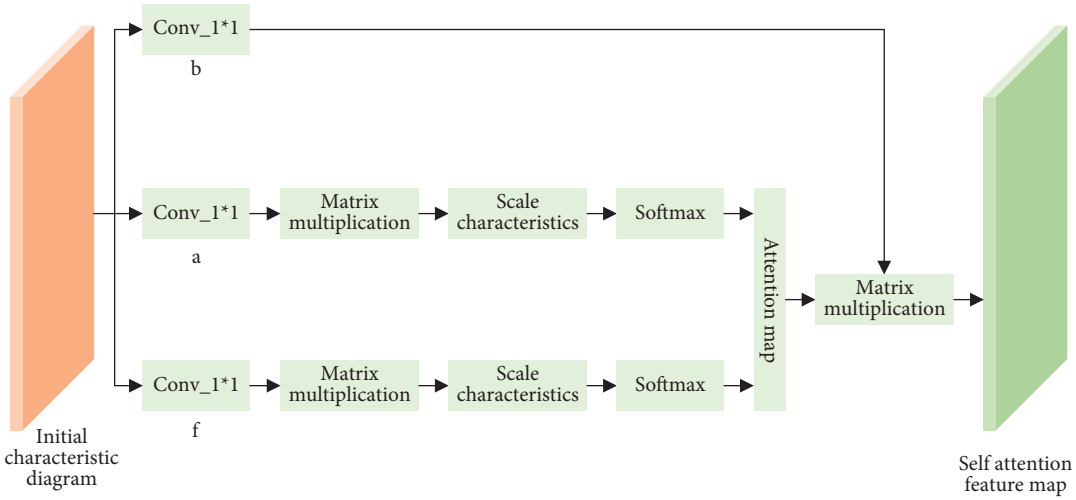


Figure 4: Self-attentive mechanism.

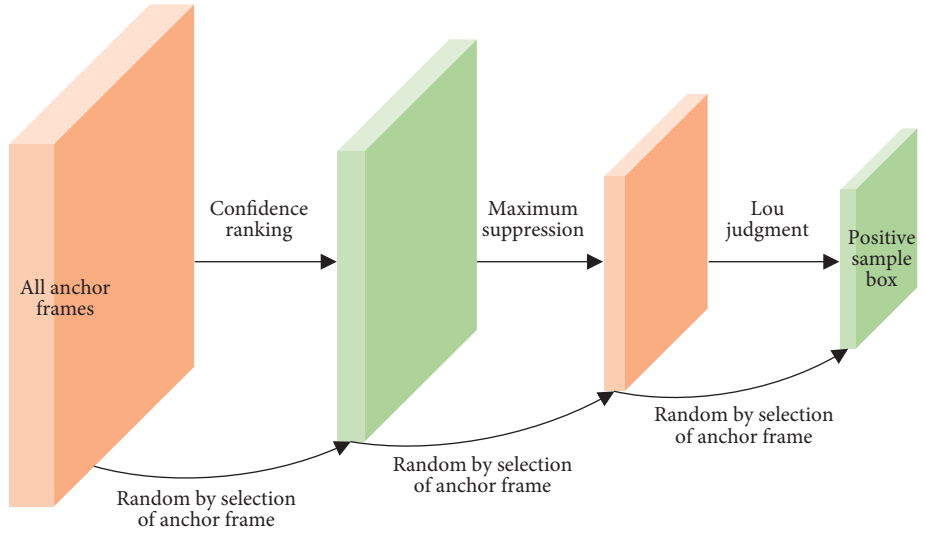


Figure 5: Anchor box enhancement.

Table 2: Experimental environment configuration.

Configuration items	Model
Programming languages	Python
Deep learning framework	PyTorch
Operating system	Ubuntu 20.04
GPU	NVIDIA GeForce RTX 3080 (2 pcs)
CPU	Intel core i9-10900K
Operating memory	64 GB
CUDA	11.4

Table 3: Performance comparison of different feature extraction networks in AI-TOD data set.

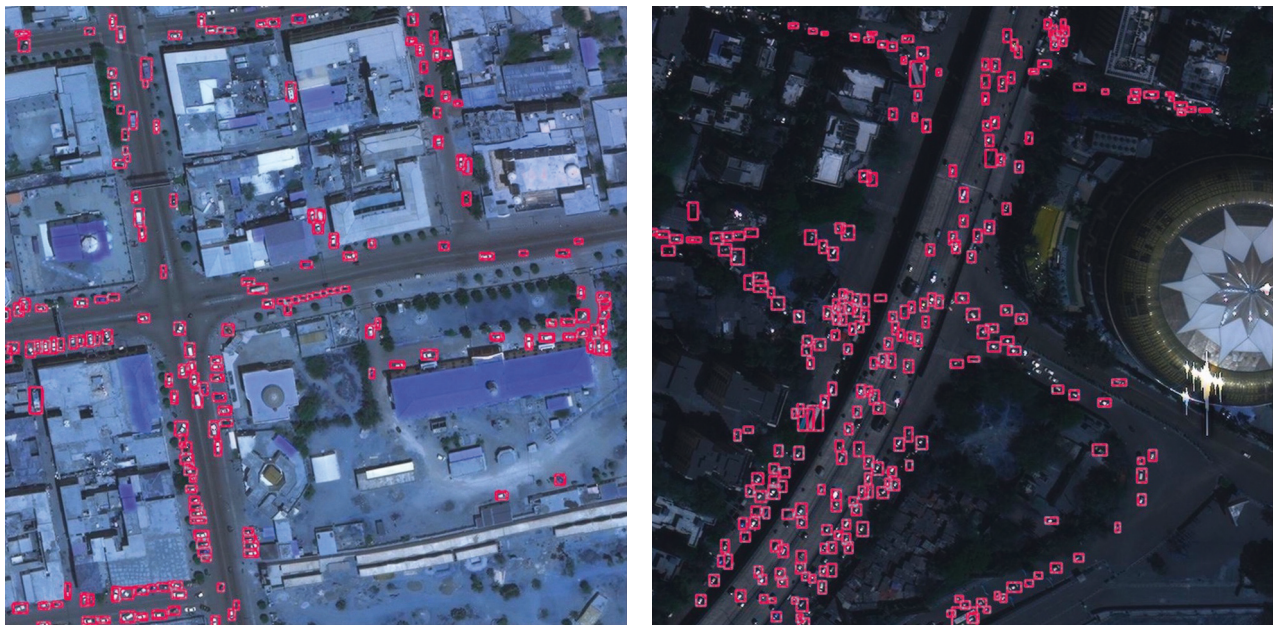
Network name	mAP (%)	AP_{50} (%)	AP_{75} (%)	AP_s (%)	AP_m (%)
Literature [23]	7.6	19.1	4.2	6.7	27.3
Literature [24]	7.1	18.5	4.5	6.5	26.5
Literature [25]	6.6	17.3	3.7	6.1	25.5
Literature [26]	6.4	15.5	3.3	5.5	23.9
Literature [27]	6.8	17.4	4.2	6.4	26.2
Proposed	8.4	19.3	5.5	7.2	20.4

feature extraction network structure with similar depth. The same data enhancement method is used for different feature extraction networks, and the FPN structure is applied after the feature extraction network. The detector and loss function adopt Generalized Focal Loss. The experimental results are shown in Table 3.

It can be seen from Table 3 that the proposed network structure has the best effect compared with other commonly used models, reaching 8.4%. Literature [23] network has a high width. Literature [24] realizes the integration of shallow information and deep information. Therefore, the two models are superior, which are 7.6% and 7.1%, respectively. The network structure proposed in this paper has both the above two characteristics. The mAP value is 1.3% higher than that in literature [24] and 0.8% higher than that in

Table 4: Analysis of ablation experiment results.

a	b	c	mAP (%)	AP_{50} (%)	AP_{75} (%)	AP_s (%)	AP_m (%)
√			5.51417	13.42249	3.45067	4.59635	10.99573
	√		0.01096	0.03062	0.00966	0.00613	0.04962
		√	0.00005	0.00024	0.00000	0.00009	0.00003
√	√		8.03065	19.13708	5.12697	6.90247	19.34725
√		√	5.70067	13.98943	3.59441	4.72061	12.61058
	√	√	0.00198	0.00442	0.00197	0.00134	0.00293
√	√	√	8.31967	19.71287	5.48544	7.53426	20.61917



(a) (b)

Figure 6: Visual result comparison.

literature [23]. Other indicators' performance is also excellent, but there is a lack of performance in the detection index APm of medium targets, which is only 20.4%. The comparative analysis of the experimental results shows that the performance of the proposed network structure in the small target detection task is better than the current mainstream feature extraction network and has good feature extraction performance. The feature fusion factor can also be adjusted according to the different application scenarios to adapt to the detection of targets with different scales.

4.3.2. Comparison of Branch Output Characteristics of Different Scales. In order to improve the detection ability of the model, the ablation experiment was set to explore the influence of the characteristic map's output by branches of different scales on the detection results. The experimental method is as follows: use AI-TOD data set as the verification set for testing, only one or two branch outputs are retained each time, and the rest are set to 0. Assuming that the output characteristic diagram with the size reduced by 4 times is a, 8 times is B and 16 times is C, the comparison results of ablation experiment are shown in Table 4.

The experimental results show that the output characteristic map of low-resolution branches has little contribution to the final detection results. After setting the feature map F3 to 0, the detection accuracy is only reduced by 0.3%, and the results of simultaneous output of F1 feature map and F3 feature map are only 0.2% higher than that of F1 single output. The experiment gives an idea for the further optimization of the model, that is, the detection speed can be improved at the minimum cost of accuracy by reducing the parameters of high-resolution branches.

4.3.3. Visualization Results. To visually verify the effectiveness of the model in this paper, compare the model in this paper without attention mechanism with the model in literature [23], and detect an aerial photograph containing dense small targets. Figure 6(a) shows the visualization results of the model in literature [23], and Figure 6(b) shows the visualization results of the model in this paper.

It can be seen from the figure that the model proposed in this paper has a stronger detection ability for small targets than the model in literature [23] and can detect more small targets.

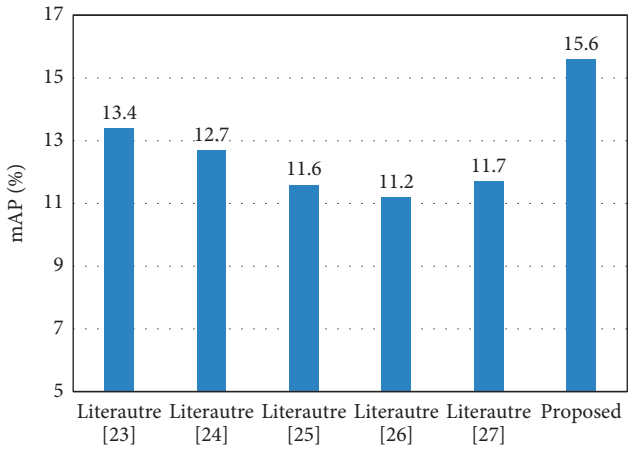


Figure 7: The comparison results of different models.

In order to further verify the value of this model, this model is compared with literature [23], literature [24], literature [25], literature [26], and literature [27]. The experimental results are shown in Figure 7.

The comparison results show that the proposed algorithm has a higher map value and stronger detection ability for small targets.

5. Conclusion

This paper studies some prominent problems of target detection and proposes a lightweight target detection model of deep learning target detection. VoVNet combined with residual structure is used as the backbone network. By integrating a self-attention mechanism, anchor design based on prior information, and selection to enhance IOU optimization, identification and classification can be realized to the maximum extent and detection ability can be enhanced. Compared with other classical algorithms, the algorithm integrating self-attention mechanism and improved anchor box strategy has stronger detection ability and higher average detection accuracy than other algorithms. Furthermore, improving the detection accuracy and the recognition ability of occluded targets is the focus of the next study and research and strives to obtain a more accurate and efficient target detection algorithm. In future work, we will optimize the model under the guidance of the research results of this paper, realize the lightweight processing of the model, improve the detection speed and the number of parameters occupied, and make the model easier to deploy on the mobile terminal.

Data Availability

The labeled data set used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The author declare that there are no conflicts of interest.

Acknowledgments

This study was funded by (1) Research on Key Technologies of Industrial Robots' Vision System Based on Deep Learning (no. 222102220120), supported by the Science and Technology Department of Henan Province, 2022; (2) Research on Motion Planning of the Six-Axis Industrial Robots Based on ROS (no. 22B413002), supported by the Education Department of Henan Province, 2022; (3) Research on the Application of the High-Efficient Data Collection System of Wireless Sensor Network Based on RF Power Supply no. 22B413003), supported by the Education Department of Henan Province, 2022; and (4) Research on Optimization Technology of Machine Learning Algorithm in Big Data Mining (no. 222102210252), supported by the Science and Technology Department of Henan Province, 2022.

References

[1] H. Liu, M. Wang, L. Liu, J. Wu, and H. Huang, "Review of small object detection based on deep learning," *Computer Engineering and Science*, vol. 43, no. 8, pp. 1429–1442, 2021.

[2] K. Li, X. Wang, H. Lin et al., "Survey of one-stage small object detection methods in deep learning," *Journal of Frontiers of Computer Science and Technology*, vol. 16, no. 1, pp. 41–58, 2022.

[3] A. Kumar, Z. J. Zhang, and H. Lyu, "Object detection in real time based on improved single shot multibox detector algorithm," *EURASIP Journal on Wireless Communications and Networking*, vol. 1, pp. 1–18, 2020.

[4] X. Zhang, K. Zhu, G. Chen et al., "Geospatial object detection on high resolution remote sensing imagery based on double multiscale feature pyramid network," *Remote Sensing*, vol. 11, no. 7, p. 755, 2019.

[5] L. Zhou, X. Rao, Y. Li, X. Zuo, B. Qiao, and Y. Lin, "A lightweight object detection method in aerial images based on dense feature fusion path aggregation network," *ISPRS International Journal of Geo-Information*, vol. 11, no. 3, p. 189, 2022.

[6] Y. Chen and H. Shin, "Multispectral image fusion based pedestrian detection using a multilayer fused deconvolutional single-shot detector," *Journal of the Optical Society of America*, vol. 37, no. 5, pp. 768–779, 2020.

[7] Y. Xiao, Z. Tian, J. Yu et al., "A review of object detection based on deep learning," *Multimedia Tools and Applications*, vol. 79, no. 33-34, pp. 23729–23791, 2020.

[8] L. Zhu, X. Geng, Z. Li, and C. Liu, "Improving YOLOv5 with attention mechanism for detecting boulders from planetary images," *Remote Sensing*, vol. 13, no. 18, p. 3776, 2021.

[9] B. Jiang, R. K. Qu, Y. D. Li, and C. L. Li, "Object detection in UAV imagery based on deep learning: review," *Acta Aeronautica et Astronautica Sinica*, vol. 42, no. 4, pp. 137–151, 2021.

[10] S. Q. Ren, K. M. 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.

[11] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object detection with deep learning: a review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[12] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep learning methods for small

object detection," *Expert Systems with Applications*, vol. 172, no. 4, Article ID 114602, 2021.

[13] H. Liang, Q. Wang, Q. Zhang, and C. X. Li, "Small object detection technology: a review," *Computer Engineering and Applications*, vol. 57, no. 1, pp. 17–28, 2021.

[14] Y. Liu, H. Liu, J. Fan et al., "A survey of research and application of small object detection based on deep learning," *Acta Electronica Sinica*, vol. 48, no. 3, pp. 590–601, 2019.

[15] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: a large data set for nonparametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[16] S. Han, B. Zhang, and L. I. Wei, "Small target detection in airport scene via modified faster -RCNN," *Journal of Nanjing University of Aeronautics & Astronautics*, vol. 51, no. 6, pp. 735–741, 2019.

[17] X. Zeng, W. Ouyang, J. Yan et al., "Crafting GBD-net for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 9, pp. 2109–2123, 2018.

[18] C. Zheng, Y. Zhang, and H. Hu, "Object detection enhanced context model," *Journal of Zhejiang University*, vol. 54, no. 3, pp. 529–539, 2020.

[19] R. Y. Zhang, X. J. Jiang, J. S. An, and T. S. Cui, "Design of global-contextual detection model for optical remote sensing targets," *Chinese Optics*, vol. 13, no. 6, pp. 1302–1313, 2020.

[20] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1331–1344, 2021.

[21] J. Yan, L. Zhao, W Diao, H. Wang, and X. Sun, "AF-EMS detector: improve the multiscale detection performance of the anchor-free detector," *Remote Sensing*, vol. 13, no. 2, p. 160, 2021.

[22] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "WiderPerson: a diverse dataset for dense pedestrian detection in the wild," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 380–393, 2020.

[23] J. Sokolić, R. Giryes, G. Sapiro, and M. R. D. Rodrigues, "Robust large margin deep neural networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4265–4280, 2017.

[24] J. Ji, S. Li, J. Xiong, P. Chen, and Q. Miao, "Semantic image segmentation with propagating deep aggregation," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9732–9742, 2020.

[25] S. Li, M. Deng, J. Lee, A. Sinha, and G. Barbastathis, "Imaging through glass diffusers using densely connected convolutional networks," *Optica*, vol. 5, no. 7, pp. 803–813, 2018.

[26] J. Zhang, C. Lu, X. Li, H. J. Kim, and J. Wang, "A full convolutional network based on DenseNet for remote sensing scene classification," *Mathematical Biosciences and Engineering*, vol. 16, no. 5, pp. 3345–3367, 2019.

[27] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-Object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network," *Remote Sensing*, vol. 12, no. 9, p. 1432, 2020.