Hindawi
Advances in Multimedia
Volume 2022, Article ID 2953700, 9 pages
https://doi.org/10.1155/2022/2953700

Research Article

A Small Target Detection Method Based on the Improved FCN Model

Guofeng Ma [1,2]

[1]*School of Artificial Intelligence, Zhengzhou Railway Vocational and Technical College, Zhengzhou 450052, China*
[2]*Henan Provincial Research Center of Wisdom Education, Zhengzhou 451460, China*

Correspondence should be addressed to Guofeng Ma; 10857@zzrvtc.edu.cn

Received 26 May 2022; Revised 9 July 2022; Accepted 22 July 2022; Published 5 September 2022

Academic Editor: Qiangyi Li

Traditional object detection is mainly aimed at large objects in images. The main function achieved is to identify the shape, color, and trajectory of the target. However, in practice, small target objects in the image must be detected in addition to large targets. Common small target detection (STD) is mainly used in intelligent transportation, video surveillance, and other fields. Small targets have small size, few pixels, and low resolution and are easily blocked. Small object detection has emerged as a research problem and a hotspot in the field of object detection. This study proposes an improved FCN model based on the full convolutional neural network (FCN) and applies it to the STD. The following is the central concept of the proposed method. Small targets are prone to occlusion and deformation in the image data. The deformation here is mainly reflected in the larger shape obtained by shooting at different angles. Therefore, it is a challenge to fully and accurately obtain the characteristics of small targets. The traditional method based on multilayer feature fusion cannot achieve ideal results for STD. This study is based on FCN and introduces a spatial transformation network. The network can optimize the handling of problems such as partial occlusion or deformation of small objects. The use of a spatial transformation network can alleviate the problem of poor feature extraction caused by partial occlusion or deformation of small targets, improving final detection accuracy. The experimental results on public datasets show that the proposed method outperforms other deep learning algorithms (DLA) in the detection accuracy of small target objects. Furthermore, the model's training time is reduced. This study's research provides a good starting point for the detection, recognition, and tracking of some small objects.

1. Introduction

STD technology is widely used in real-life production, such as intelligent transportation [1–3], medical image diagnosis in medicine [4–6], image retrieval [7–9], remote sensing image analysis [10–12], and military applications [13–15]. Target detection is mainly used to identify the target object from the data to be detected, including the position, shape, size, and color of the object. Traditional object detection-related research mainly focuses on large objects. In recent years, people have begun to focus on STDs. The detection of small objects is more difficult than the detection of large objects. For example, in unmanned scenarios, vehicles need to avoid in time, and small objects also need to be correctly identified [16]. In medical image detection, when small tumors appear in the body, STD can assist doctors in making a diagnosis [17]. In remote sensing images, there are many elements, such as buildings, bridges, vehicles, and so on. The identification of these contents is of great significance [18]. STD is very difficult because the size and resolution of the data obtained are low. The accuracy of STD has gradually improved as computer vision technology has matured. On the other hand, the need for STD in various application scenarios also increases simultaneously. Therefore, it becomes very urgent to find a general STD method with a good detection effect and insensitive to data.

Target detection mainly includes two categories, detecting specific target objects and detecting specific categories of target objects. The so-called specific target objects usually refer to specific target object instances such as landmarks or faces. The specific categories of objects refer to different categories of target objects as research objects, such as people, cars, bicycles, and other different categories of target objects. Initially, people

only focused on the detection of a specific category or categories of objects. Later, with the gradual improvement of target detection technology and the continuous in-depth research of scholars, a single category of target detection can no longer meet people's needs. Therefore, the construction of complex, complete, and general target detection models has become the direction of scholars' continuous in-depth research and exploration. The core idea of target detection based on machine learning is to input an image and calibrate the area where the target object is located. Feature extraction for the target object. The extracted features are input into the trained classifier for object recognition. References [19–21] are typical machine learning-based object detection-related research. The problem with this method is that the hand-designed features have a poor generalization and poor robustness, and cannot cope with complex detection tasks well. Therefore, STD based on DLA [22] were born one after another. Reference [23] introduces R–CNN to object detection. The core idea of this method is to extract multiple candidate frames after inputting image data. The candidate regions are input into a convolutional neural network (CNN) to obtain candidate features. Finally, the candidate features are input to the support vector machine (SVM) for target object recognition. Reference [24] is based on reference [23], and the STD results obtained are significantly improved. Reference [25] proposes an improved CNN, the Faster R–CNN algorithm.

Through comparative analysis of the above studies, it is found that although the success rate of STD is getting higher and higher, there are still some difficulties that need to be solved. In the detection process based on deep learning, there are still the following problems. First, the features extracted by each convolutional layer in the model are different. After the convolutional layer extracts the features, they are fed to the prediction layer. Since the proportion of small objects in the entire image is very small, the information is more dispersed after feature extraction through multiple convolutional layers, which eventually leads to the model being unable to accurately predict the type or location of small objects. Second, due to the small proportion of small objects in the image data, the proportion of positive and negative samples is unbalanced, which easily leads to the overfitting of the trained model. Typically, the number of negative samples far exceeds the number of positive samples. Third, most models improve detection rates by deepening the network. Although this method can upgrade the detection accuracy of the model, the large model will lead to a decrease in the detection rate. To address the aforementioned issues, this study combines Spatial Transformation Network (STN) and FCN. The proposed network spatially transforms and aligns the input image and output feature maps, respectively, during training. Thus, the problem of difficult feature extraction and poor feature learning effect caused by the angle of the shooting of small objects is improved.

2. Knowledge about STD

2.1. Traditional STD. Dimensions are further divided into relative dimension definitions and absolute dimension definitions. The relative size is calculated based on the width and height of the original image. Objects less than or equal to one-tenth the width or height of the original image are considered small objects. Absolute dimensions are based on standards established by the international organization SPIE. In $N \times N$ images, objects smaller than 0.12% of the overall image are identified as small objects. The traditional target detection algorithm uses a preset window to perform sliding traversal on the input image to extract candidate frames. The obvious flaw of this method is that it does not capture the unique characteristics of the target. Methods are inefficient and difficult to execute. Figure 1 depicts the flow of the traditional target detection method.

2.2. STD Based on DLA. The hand-designed features in traditional target detection algorithms have the problems of weak generalization and poor robustness. Therefore, traditional methods cannot efficiently deal with complex detection tasks. In this context, STD based on DLA emerges as the times require. DLA can automatically extract features. The extracted feature information is rich, which can effectively improve the precision of STD. The progress of STD based on DLA is shown in Figure 2.

The two-stage algorithm shown in Figure 2 is the most widely studied. The structure of the two-stage class algorithm is shown in Figure 3. After extracting the sample features, the training is divided into two steps. First, while training the network, the feature map will be classified into regions, so as to make a preliminary prediction of the target location. Second, perform precise positioning and correction for the predicted position.

In addition to the two-stage algorithm, some scholars have also proposed STD based on the single-stage algorithm. Compared with the Two-stage algorithm, the single-stage algorithm mainly gives the category and location information of the small target through the backbone network. Although this method sacrifices the accuracy of detection, it improves the training speed of the algorithm. Figure 4 shows the structure of the Single-stage class algorithm. First, the image to be detected is input to the convolutional layer. Second, directly perform probabilistic detection on the category of the object and locate the area where the object is located. The Single-stage algorithm abandons the candidate region generation stage when performing STD, and only needs a single detection to obtain the detection result.

3. Improved FCN Model

3.1. R-FCN. In traditional convolutional neural networks, due to the excessive use of pooling layers, the translation-invariance performance of the network is improved when performing image classification tasks. However, translation variance is required when performing object detection tasks, requiring the model to be position-sensitive. In order to solve the above problems, R-FCN specially encodes the position information, and additionally outputs a position-sensitive score map after the traditional full convolution layer, so as to maintain the entire frame volume. The structure of the product also realizes the "translational

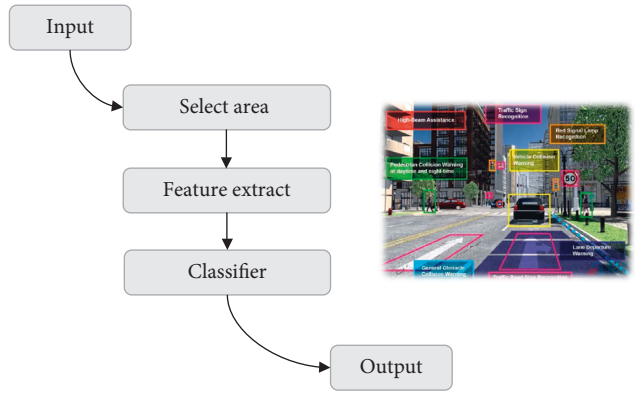

Figure 1: Traditional target detection process.

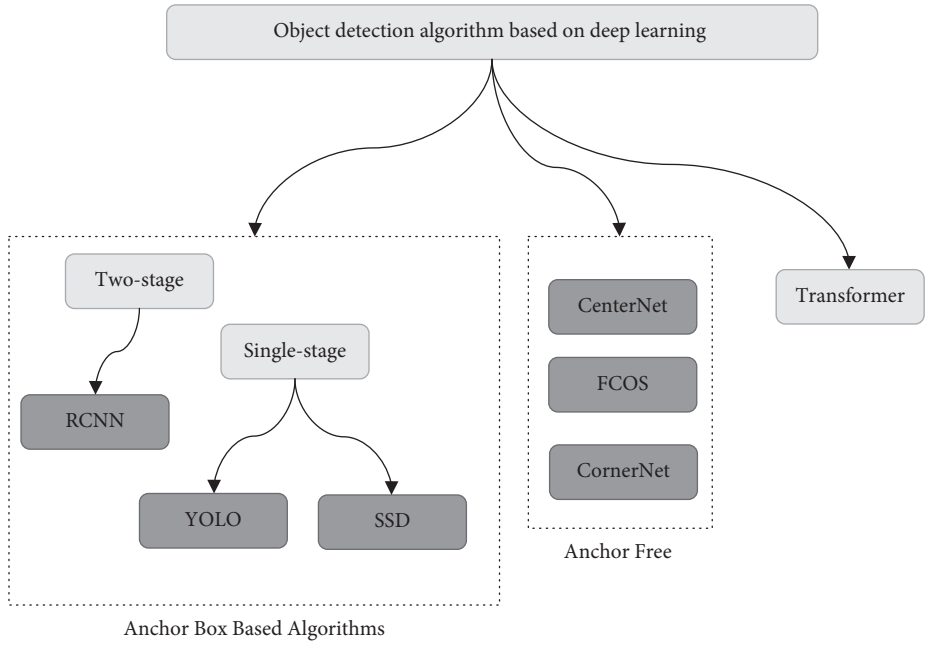

Figure 2: Research progress of STD based on DLA.

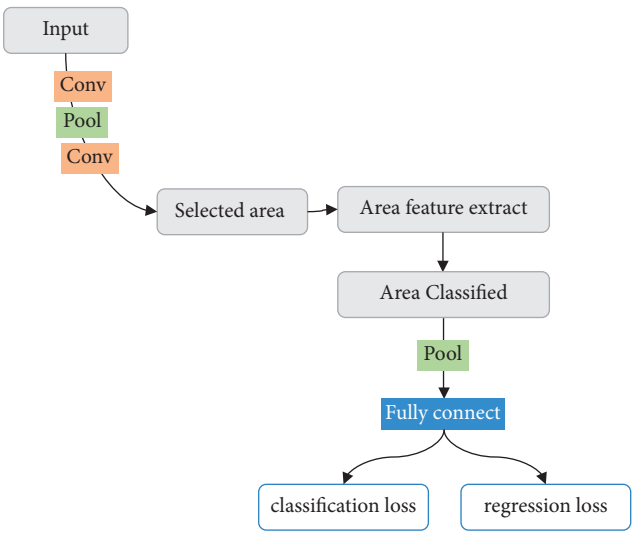

Figure 3: Two-stage class algorithm structure diagram.

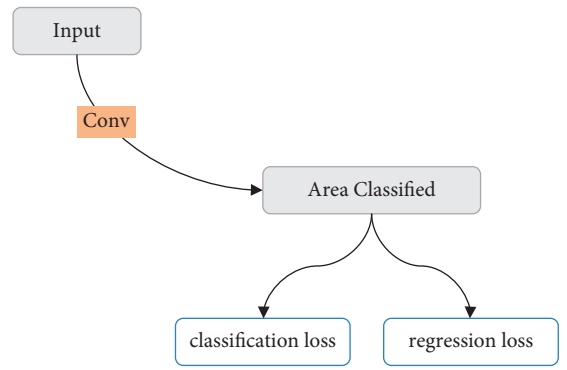

Figure 4: Single-stage class algorithm structure diagram.

changeability". To emphasize position sensitivity, R-FCN adds a position-sensitive RoI pooling layer on top of the entire fully convolutional network. In this way, the entire fully convolutional layer can both share computation and encode positions.

R-FCN includes an object detection module, which has a significant effect on the localization and classification of object detection. The network training process mainly includes two stages: region proposal and region classification. The purpose of the region proposal stage is to select candidate regions. The region classification stage performs classifier operations on candidate regions. Figure 5 shows the R-FCN structure. As can be seen from the figure, R-FCN mainly consists of 4 parts, namely the basic convolutional network, RPN network, ROI Pooling pooling layer, and classification network.

The basic network selected by R-FCN is a deep residual network with a depth of 50. R-FCN mainly performs convolution and pooling on the input image data. The region proposal part uses the ResNet-50 network. The output of this network is the candidate region. The region proposal network uses multiple convolution kernels to convolve the input samples to obtain a tensor. This tensor is fed into two separate convolutional layers.

Figure 6 presents the structure of the region proposal model. The principle of this network work is to use multiple search boxes of different sizes to search for a certain area. The unique name of the search box is the anchor. For each image input to the RPN network, there are about 20,000 initial search boxes. After subsequent de-overlapping processing, the number of search boxes will be greatly reduced.

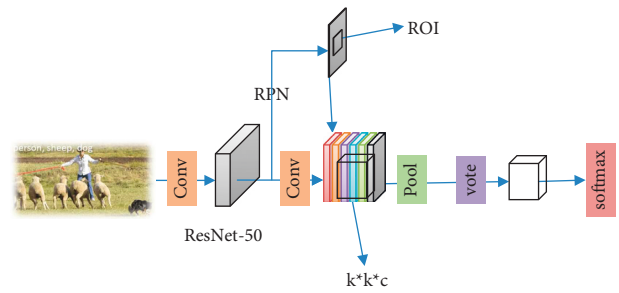

Figure 5: R-FCN structure.

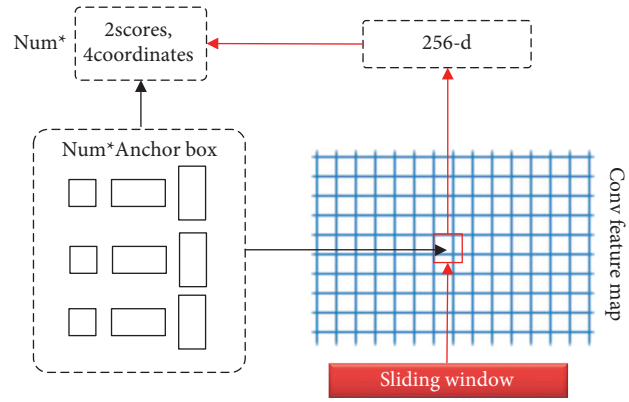

Figure 6: Regional proposal network structure diagram.

3.2. Improved FCN. When the traditional fully convolutional neural network detects small objects, the region suggestion operation is performed in each iteration. Moreover, thousands of candidate boxes often need to be extracted. Therefore, the model is very time-consuming to train. R-FCN improves the region proposal process, and the number of candidate boxes extracted is reduced from thousands to hundreds. However, R-FCN has not achieved the desired effect in terms of detection accuracy. In order to improve the R-FCN model and improve its detection accuracy for small targets. When improving the model, the following rules need to be followed. If the number of convolutional layers is large, the recall rate of the model is high, but the localization ability of the target position is poor. With a small number of convolutional layers, although the localization ability of the target position can be enhanced, the recall rate will be reduced. Therefore, a good model needs to grasp the relationship between the two.

The new hyperfeature designed in this study can well achieve this purpose. The core idea is to extract the feature map and give the location and other information of the small target as accurately and efficiently as possible. This study combines hyperfeatures and R-FCN to propose an improved FCN model. The model in this study changes the network structure and shortens the time required for network training. Hyperfeatures are extracted through the HyperNet model. The frame structure of HyperNet is shown in Figure 7. First, the image to be tested is input to the convolution layer to extract feature data. Second, the feature data extracted from each layer are fused together and input into a unified space to generate hyperfeatures. Third, 160 candidate

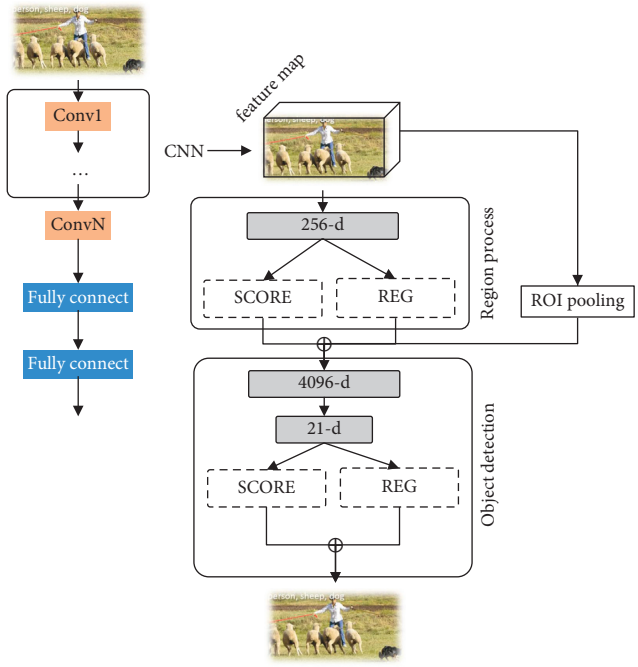

Figure 7: Principle of HyperNet.

regions are extracted using a region proposal network. Finally, we use the detection model to classify the candidate regions to obtain the final detection result.

The improved FCN mainly includes four steps of hyperfeature extraction, region proposal generation, target detection, and fusion training. Details of each step are described below:

The first is hyperfeature extraction. For different convolutional layers, HyperNet gives different processing methods. For the bottom convolutional layer, it is mainly to increase the maximum pooling layer. For high-level convolutional layers, deconvolution operations are added. The essence of this operation is upsampling. Each convolutional layer corresponds to a sampled result. The function of convolution is not only to extract the information of the image, but also to realize processing such as feature compression. The compressed features also need to be normalized to obtain feature cubes. This feature cube is the hyperfeature. Typically, the normalization method used in the compressed space is local response normalization.

The second is the generation of region proposals. The depth of the network is crucial to the feature extraction ability of the network. Therefore, a small convolutional neural network is added to the selection of candidate regions section. This network has very few layers, only three layers. They are RoI Pooling, convolutional layer, and fully connected layer, respectively. Connected behind the network are two output layers. After passing through this small convolutional neural network, each image will generate tens of thousands of initial candidate boxes. There are some errors in the size and resolution of these candidate boxes, and the next layer of the network will select and improve these initial candidate boxes.

For the output bins, the ROI pooling layer usually uses a dynamic max pooling method for processing. To optimize the network, two additional layers are added after the ROI Pooling layer of the HyperNet network. These two layers contain encoding and scoring layer, respectively. The encoding layer is responsible for encoding the position of the candidate box output by the previous layer to generate a three-dimensional feature cube. And encode the feature cube again to generate a 256-dimensional feature vector. The scoring layer is responsible for outputting the probability of the existence of small targets. When the scores of all candidate boxes are calculated, it usually happens that the scores are the same. To minimize this occurrence, the non-maximum suppression (NMS) method is used. The idea of this method is to decide whether to retain each candidate box by setting a threshold. Determine the relationship between the IoU threshold of all candidate frames and the set threshold. When the IoU threshold is greater than or equal to the set threshold, the candidate frame is retained. When the IoU threshold is less than the set threshold, the candidate frame is removed. In most studies, the loU threshold is usually set to 0.7. Before NMS operation, more than 1000 candidates are usually generated per sample. After the NMS operation, the number of candidate regions is reduced by 80 percent.

The third is target detection. Existing research has proved that in the network structure, the method based on the cross-connection of the fully connected layer and the dropout layer has the potential to improve model performance. Based on the inspiration of this theory, this study improves the HyperNet structure in two aspects. One is to add a convolutional layer before FC to improve the classification effect. The size of the convolutional layer is set to $3 \times 3 \times 63$. The introduction of convolutional layers can also reduce the feature dimension and optimize subsequent calculations. The second is to change the dropout ratio from 0.5 to 0.3. Experiments show that this operation can improve the training efficiency of the network.

In the candidate region extraction process, the network will have two outputs. All candidate boxes can calculate $M+1$ values and adjustment values of $4 \times M$ candidate boxes. M indicates the number of target categories contained in the image to be tested. The 1 in $M+1$ represents the background of the image. The introduction of the NMS method can reduce the existence of identical candidate boxes. But the number of candidate boxes removed would not be many. Most of the candidate boxes have been removed in the previous step.

The fourth is fusion training. The candidate boxes are classified by setting a threshold. When the candidate frame's IoU threshold is greater than or equal to the set threshold, the candidate frame is classified as a positive sample. When a candidate box's IoU threshold is less than the set threshold, the candidate box is classified as a negative sample. When training the network, we choose any deep network as the base network. In this study, a deep neural network was selected for training. The base network is used to separately train the HyperNet network to generate hyperfeatures. When the candidate frame of the image to be inspected is extracted, the target detection module will classify the candidate frame and continue to call back to train a better HyperNet. When HyperNet is trained, hyperfeatures are generated. This feature is mainly used for final classification. During the training of the improved FCN model proposed in this study, the extraction of candidate regions and the detection network are both trained separately. After joint training, information is shared between the two networks to extract hyperfeatures. Finally, the two networks are merged into a completely large network.

4. Experimental Setup and Result Analysis

4.1. Experimental Data. This study uses two public datasets for experimental analysis, namely, PASCAL VOC and Microsoft Common Objects in Context (COCO) datasets. The details of the two datasets are as follows:

4.1.1. PASCAL VOC. PASCAL VOC mainly includes two subdatasets, PASCAL VOC207 and PASCAL VOC 2012. The PASCAL VOC data have a total of 20 categories, as shown in Table 1.

Among them, the number of images and objects of PASCAL VOC 2007 and PASCAL VOC 2012 are shown in Table 2. The data in the table show that the database contains a large number of samples as well as a large number of small target samples. It serves as the test library for the majority of the current STD algorithms.

4.1.2. COCO. COCO mainly includes two subdatasets: COCO-2014 and COCO-2017. The number of images in the training set and test set in the COCO-2014 dataset is 82783

Table 1: Data object categories.

Category	Target
Humanity	People
Animal	cat, cow, horse, bird, dog, sheep
Indoor	Sofa, chair, bottle, dining table, plant, TV
Vehicle	Bicycles, planes, boats, cars, trains, buses, motorcycles

Table 2: Dataset details.

Dataset	Training set		Validation set		Test set	
	Picture	Object	Picture	Object	Picture	Object
PASCAL VOC 2007	2501	6301	2510	6307	4952	12032
PASCAL VOC 2012	5717	13609	5823	13841	—	—

and 118287, respectively. There are a total of 80 object classes in this dataset, which are divided into large, medium, and small objects. The size and proportion of each target are shown in Table 3.

4.2. Evaluation Indicators and Experimental Environment. The quality of a model is usually evaluated from three aspects, namely speed, stability, and Mean Average Precision (mAP). The calculation of the mean precision means is IoU dependent. A common IoU confidence threshold for object detection algorithms is 0.5. If the loU is greater than or equal to 0.5, it is considered that the box contains objects, which is a correct detection. If the loU is less than 0.5, it is considered that the box does not contain objects, which is a wrong detection. Given class C, the precision of object detection in an image is calculated as follows:

$$P_i = \frac{T_i}{N_i}, \tag{1}$$

where P_i is the detection accuracy of i-category objects, T_i is the number of i-type objects in the detected image, and N_i is the total number of i-type objects in the image. On the premise that the accuracy of each image is known, the average accuracy of class i samples in the test set can be calculated:

$$AP_i = \frac{\sum T_i}{N_i}. \tag{2}$$

In (2) means that the average precision value of the i category is equal to the sum of the i category precisions of all pictures in the test set and is higher than the number of all pictures in the test set containing the i category objects. When the detection accuracy value of each class is calculated, the global average loss precision can be calculated. N_{classes} represents the total number of classes.

$$\text{mAP} = \frac{\sum AP_i}{N_{\text{classes}}}. \tag{3}$$

Precision alone is not enough to measure the performance of a model. The precision and recall curve can not only characterize the detection accuracy of the model, but also characterize the robustness and stability of the model. (4) is the calculation formula of recall rate (R).

Table 3: COCO-2014 dataset.

Category	Proportion (%)	Min pixel	Max pixel
Small target	41	0×0	32×32
Medium target	34	32×32	96×96
Big target	25	96×96	$\infty \times \infty$

$$R = \frac{N_s^+}{N_{\text{all}}^+}, \tag{4}$$

where N_s^+ denotes the number of positive samples that were correctly predicted. N_{all}^+ denotes the total number of positive samples.

The environment used in this experiment is: Ubuntu 20.10 64 bit operating system, Intel(R) Core(TM)i7-7700 processor, 64 GB memory, NVIDIA GeForce GTX 3060Ti graphics card, and the experimental model is studied in the deep learning framework TensorFlow 2.3.0. The number of network iterations is set to 5000 during model training, and the learning rate is set to 0.001.

4.3. Analysis of Experimental Results. The research in this study belongs to STD based on DLA. For the fairness of the experiment, the following DLAs are introduced as a comparative study. The main comparison algorithms are CNN [26], RNN [27], FCN [28], HR-FCN [29], Fast R-CNN [30], YOLOv3 [31], and STDN [32]. The detection precision of eight small objects in the 2007 dataset is shown in Table 4.

Each algorithm in Table 4 has a better detection effect on the four target objects of Boat, TV, Bird, and Sheep. This shows that these kinds of objects have rich and obvious features compared to other small objects. For Bottle, comparing the detection results of different algorithms, it can be seen that YOLOv3 has the highest detection rate, followed by Fast R-CNN, and the third is the proposed algorithm. For plant, the detection rates of all algorithms are relatively low, Fast R-CNN has the best detection effect, followed by the proposed algorithm but the gap between the two is smaller. From Chair to MAP, the detection results of the proposed method are higher than other methods. Therefore, on the whole, the detection effect of the proposed algorithm is better.

Figure 8 presents the average P-R curves of each model for eight small object detections on the VOC 2007 dataset. The larger the area of the curve, the better the model's performance. Figure 8 shows the curve area obtained by each model very intuitively. The largest area is table, and boat, which shows that the detection effect of these two small targets is good. In addition, we observed that no matter what kind of small target results, the P-R curve area obtained by the model used in this paper is the largest compared with other models, which shows that the method in this paper is the best.

Different IOU settings will result in different mAPs. For the COCO dataset, different confidence levels have different

Table 4: mAP of each model on PASCAL VOC 2007.

Target\method	CNN	RNN	FCN	HR-FCN	Fast R-CNN	YOLOv3	STDN	Proposed
Bottle	52.25	54.03	56.61	57.86	62.43	65.84	54.71	61.96
Plant	41.87	49.91	44.06	51.00	57.56	47.74	53.79	56.27
Chair	57.87	48.59	59.33	50.52	54.46	55.94	53.95	62.98
Boat	64.79	69.36	59.63	64.79	72.43	73.39	69.63	73.57
Tv	74.46	75.74	72.57	69.57	77.01	72.76	68.01	77.82
Table	67.47	76.50	75.79	64.98	70.78	76.31	68.40	78.74
Bird	77.33	77.30	77.39	76.87	79.47	69.25	69.94	81.16
Sheep	74.71	69.69	73.06	70.94	78.29	77.63	75.54	82.30
MAP	63.84	65.14	64.81	63.32	69.05	67.36	64.25	71.85

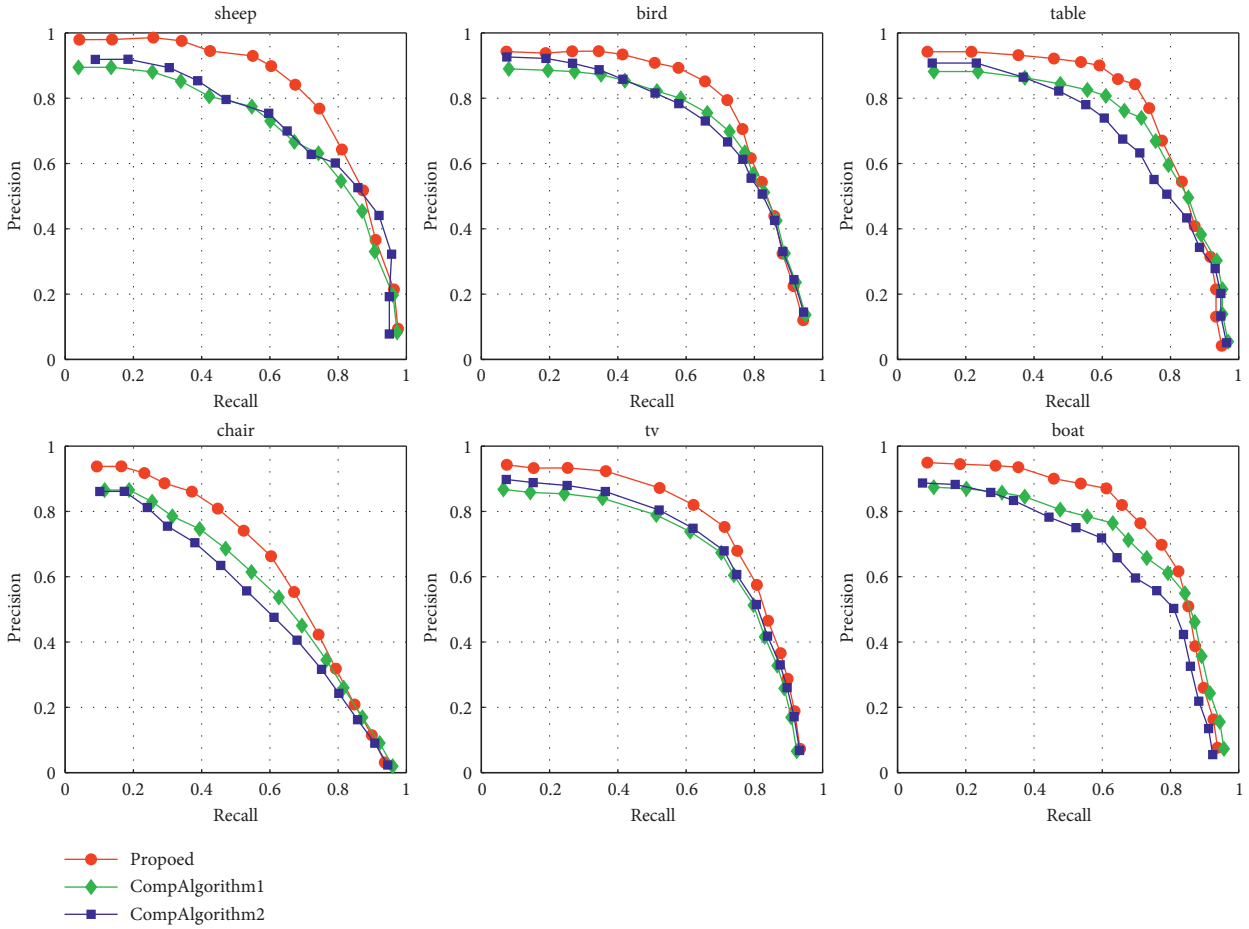

Figure 8: P-R curve on the VOC2007 dataset.

detection precision. AP_x represents the detection precision obtained when the confidence level is x, and the subscript x is from 50 to 75. Table 5 shows the AP values calculated by each model on the COCO dataset. Figure 9 graphs the AP values obtained by each model for easy comparison and viewing.

The experimental results on the COCO dataset show that the highest detection rate is obtained when the confidence level is 50. On the other hand, the AP of the proposed algorithm is optimal in most cases, no matter how the confidence changes. Followed by YOLOv3, when the confidence is 60, the AP of the two algorithms is the same. When the confidence level is 75, the AP obtained by YOLOv3 is higher than the proposed algorithm. The APs of other comparison algorithms are lower than YOLOv3 and the proposed model. The experimental results demonstrate that the proposed method can produce a more ideal AP.

Table 5: AP value on COCO dataset.

Model	AP_{50}	AP_{55}	AP_{60}	AP_{65}	AP_{70}	AP_{75}
CNN	53.22	51.22	51.09	50.31	47.07	42.12
RNN	52.19	50.57	48.33	44.05	42.81	36.91
FCN	53.53	51.44	47.91	44.68	42.57	38.19
HR-FCN	55.21	53.21	50.53	48.65	45.45	43.38
Fast R-CNN	55.16	54.63	53.86	52.91	51.52	48.02
YOLOv3	57.84	57.46	55.98	53.84	52.24	51.59
STDN	54.75	51.75	49.91	47.50	44.11	38.67
Proposed	59.97	59.46	55.53	54.93	53.47	48.19

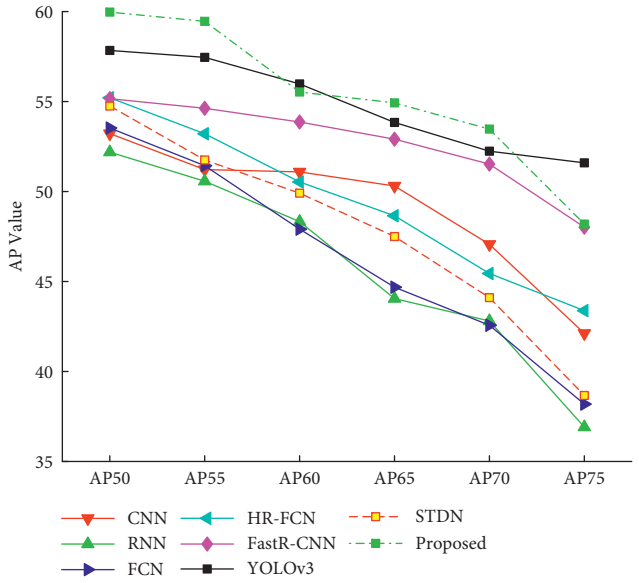

Figure 9: AP obtained by each model under different confidence levels.

5. Conclusion

With the development of technology, people have put forward higher requirements for STD in terms of speed and precision. The difficulty of STD is that the area of the target is small, so its proportion in the original image is small, and it is difficult for the detection algorithm to extract rich and effective features. The end result is that small object detection is not effective. Second, the neural network is dominated by the large target in the learning process, and the small target is ignored in the whole learning process, which leads to the poor detection effect of the small target, especially if there are many network layers, the characteristic information of the small target will be lost. In order to further improve the effect of STD, this study proposes an improved FCN model. Its core idea is to embed the spatial transformation network into the framework of FCN. By processing the deformation or multiangle input samples, it is more convenient for subsequent classification and identification. This can avoid the situation of low recognition accuracy caused by small target deformation or multiangle problems. The improved FCN mainly includes four steps of hyperfeature extraction, region proposal generation, target detection, and fusion training. The experimental results on two public datasets show that the proposed algorithm can achieve ideal detection accuracy. There are still several shortcomings in this study, such as when the background of the picture is complex, it will interfere with the detection of small objects, thus affecting the detection results. In addition, the classifier of the used network can be further optimized later to improve the classification accuracy and efficiency of the classifier.

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by Zhengzhou Railway Vocational and Technical College.

References

[1] D. Sirohi, N. Kumar, and P. S. Rana, "Convolutional neural networks for 5G-enabled Intelligent Transportation System: a systematic review," *Computer Communications*, vol. 153, pp. 459–498, 2020.

[2] A. Mhalla, T. Chateau, S. Gazzah, and N. E. B. Amara, "An embedded computer-vision system for multi-object detection in traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4006–4018, 2019.

[3] A. Kausar, A. Jamil, N. Nida, and M. H. Yousaf, "Two-wheeled vehicle detection using two-step and single-step deep learning models," *Arabian Journal for Science and Engineering*, vol. 45, no. 12, Article ID 10755, 2020.

[4] S. Ito, K. Ando, K. Kobayashi et al., "Automated detection of spinal schwannomas utilizing deep learning based on object detection from magnetic resonance imaging," *Spine*, vol. 46, no. 2, pp. 95–100, 2021.

[5] A. A. Mikhaylichenko and Y. M. Demyanenko, "Detection of the bone contours of the knee joints on medical X-ray images," *Computer Optics*, vol. 43, no. 3, pp. 455–463, 2019.

[6] A. Sedik, H. M. Emara, A. Hamad et al., "Efficient anomaly detection from medical signals and images," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 739–767, 2019.

[7] R. Mukherjee, M. Melo, V. Filipe, A. Chalmers, and M. Bessa, "Backward compatible object detection using HDR image content," *IEEE Access*, vol. 8, Article ID 142736, 2020.

[8] A. A. Zakharov, A. E. Barinov, A. L. Zhiznyakov, and V. S. Titov, "Object detection in images with a structural descriptor based on graphs," *Computer Optics*, vol. 42, no. 2, pp. 283–290, 2018.

[9] L. V. Vishnyakova, V. Y. Kim, K. V. Obrosov, N. K. Obrosova, and A. I. Rodionov, "Search-detection-recognition: simulation via thermal images with varying quality," *Journal of Computer and Systems Sciences International*, vol. 59, no. 6, pp. 905–917, 2021.

[10] J. J. Bapu and D. J. Florinabel, "Real-time image processing method to implement object detection and classification for remote sensing images," *EARTH SCIENCE INFORMATICS*, vol. 13, no. 4, pp. 1065–1077, 2020.

[11] M. N. Raju, K. Natarajan, and C. S. Vasamsetty, "Object recognition in remote sensing images based on modified backpropagation neural network," *Traitement du Signal*, vol. 38, no. 2, pp. 451–459, 2021.

[12] S. Karim, Y. Zhang, SL. Yin, I. . Bibi, and A. A. Brohi, "A brief review and challenges of object detection in optical remote sensing imagery," *Multiagent and Grid Systems*, vol. 16, no. 3, pp. 227–243, 2020.

[13] A. A. Bayramov and E. G. Hashimov, "Assessment of invisible areas and military objects in mountainous terrain," *Defence Science Journal*, vol. 68, no. 4, pp. 343–346, 2018.

[14] M. L. Pawelczyk and M. Wojtyra, "Real world object detection dataset for quadcopter unmanned aerial vehicle detection," *IEEE Access*, vol. 8, pp. 174394–174409, 2020.

[15] M. Kowalski, "Hidden object detection and recognition in passive terahertz and mid-wavelength infrared," *Journal of Infrared, Millimeter and Terahertz Waves*, vol. 40, no. 11-12, pp. 1074–1091, 2019.

[16] S. Selim, N. K. Sonmez, M. Coslu, and I. Onur, "Semi-automatic tree detection from images of unmanned aerial vehicle using object-based image analysis method," *JOURNAL OF THE INDIAN SOCIETY OF REMOTE SENSING*, vol. 47, no. 2, pp. 193–200, 2019.

[17] E. Kot, Z. Krawczyk, K. Siwek, L. KROIicki, and P. Czwarnowski, "Deep learning-based framework for tumour detection and semantic segmentation," *Bulletin of the Polish Academy of Sciences, Technical Sciences*, vol. 69, no. 3, Article ID e136750, 2021.

[18] J. Zhang, C. M. Xie, X. Xu, Z. Shi, and B. Pan, "A contextual bidirectional enhancement method for remote sensing image object detection," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4518–4531, 2020.

[19] S. Ariffa Begum and A. Askarunisa, "Performance analysis of machine learning classification algorithms in static object detection for video surveillance applications," *Wireless Personal Communications*, vol. 115, no. 2, pp. 1291–1307, 2020.

[20] E. J. Sadgrove, G. Falzon, D. Miron, and D. W. Lamb, "Real-time object detection in agricultural/remote environments using the multiple-expert colour feature extreme learning machine (MEC-ELM)," *Computers in Industry*, vol. 98, pp. 183–191, 2018.

[21] V. K. Singh and N. Kumar, "CHELM: convex hull based extreme learning machine for salient object detection," *Multimedia Tools and Applications*, vol. 80, no. 9, Article ID 13535, 2021.

[22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, pp. 2672–2680, 2014.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, p. 1, 2015.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[25] 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.

[26] A. Chaudhuri, "Hierarchical modified Fast R-CNN for object detection," *Informatica*, vol. 45, no. 7, pp. 67–82, 2021.

[27] A. Kompella and R. V. Kulkarni, "A semi-supervised recurrent neural network for video salient object detection," *Neural Computing & Applications*, vol. 33, no. 6, pp. 2065–2083, 2020.

[28] H. Q. . Li, G. L. Cui, S. S. Guo, L. J. . Kong, and X. B. Yang, "Human target detection based on FCN for through-the-wall radar imaging," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 9, pp. 1565–1569, 2021.

[29] S. X. Wu, C. C. . Guo, and X. H. Wang, "Application of principal component analysis and adaptive median filter to improve real-time prostate capsula detection," *JOURNAL OF MEDICAL IMAGING AND HEALTH INFORMATICS*, vol. 10, no. 2, pp. 336–347, 2020.

[30] R. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[31] S. A. F. Manssor, S. Y. Sun, M. Abdalmajed, and S. Ali, "Real-time human detection in thermal infrared imaging at night using enhanced Tiny-yolov3 network," *JOURNAL OF REAL-TIME IMAGE PROCESSING*, vol. 19, no. 2, pp. 261–274, 2021.

[32] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 528–537, Salt Lake City, UT, USA, June 2018.