Hindawi
Journal of Sensors
Volume 2023, Article ID 9802607, 1 page
https://doi.org/10.1155/2023/9802607

Retraction

Retracted: Small Target Detection Algorithm Based on Transfer Learning and Deep Separable Network

Journal of Sensors

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] P. Wang, H. Wang, X. Li, L. Zhang, R. Di, and Z. Lv, "Small Target Detection Algorithm Based on Transfer Learning and Deep Separable Network," *Journal of Sensors*, vol. 2021, Article ID 9006288, 10 pages, 2021.

Hindawi
Journal of Sensors
Volume 2021, Article ID 9006288, 10 pages
https://doi.org/10.1155/2021/9006288

Research Article

Small Target Detection Algorithm Based on Transfer Learning and Deep Separable Network

Peng Wang,[1,2,3] Haiyan Wang,[1,2] Xiaoyan Li,[3] Lingling Zhang,[3] Ruohai Di,[3] and Zhigang Lv[3]

[1]*School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, Shaanxi, China 710072*
[2]*Key Laboratory of Ocean Acoustics and Sensing (Northwestern Polytechnical University), Ministry of Industry and Information Technology, Xi'an, Shaanxi, China 710072*
[3]*School of Electronic and Information Engineering, Xi'an Technological University, Xi'an 710021, China*

Correspondence should be addressed to Xiaoyan Li; lixiaoyan@xatu.edu.cn

Received 30 July 2021; Revised 31 August 2021; Accepted 7 September 2021; Published 4 October 2021

Academic Editor: Haibin Lv

With the development of deep learning, target detection from vision sensor has achieved high accuracy and efficiency. However, small target detection remains a challenge due to inadequate use of semantic information and detailed texture information of underlying features. To solve the above problems, this paper proposes a small target detection algorithm based on Mask R-CNN model which integrates transfer learning and deep separable network. Firstly, the feature pyramid fusion structure is introduced to enhance the learning effect of low-level and high-level features, especially to strengthen the information channel of low-level feature and meanwhile optimize the feature information of small target. Secondly, the ELU function is used as the activation function to solve the problem that the original activation function disappears in the negative half axis gradient. Finally, a new loss function F-Softmax combined with Focal Loss was adopted to solve the imbalance of positive and negative sample proportions. In this paper, self-made data set is used to carry out experiments, and the experimental results show that the proposed algorithm makes the detection accuracy of small targets reach 66.5%.

1. Introduction

In recent years, the aerial image target detection technology based on UAV has become one of the forefront research topics [1–3]. Due to the distance from the target, the UAV aerial images are mostly small- and medium-sized targets. From the perspective of absolute size, a small target is defined as a 32∗32 pixel target. In terms of relative size, if the target occupies 0.1 times the size of the whole picture, it can be considered as a small target [4–6]. The traditional target detection algorithm is easy to cause misdetection and missed detection of small targets in these image processing, and the detection rate of small targets is low, so small target detection is the focus and difficulty in this field [7–9].

Small target detection is a very important field in image processing, and it is only in recent years that more and more attention has been paid to the research of small target detection [10, 11]. Different Gaussian methods are using in deep learning algorithms to detect small targets in maritime infrared images. However, due to small imaging area of small targets in infrared images and insignificant target features, traditional Gaussian methods have problems such as high false positive rate in target detection [12–15]. A general band selection algorithm based on high-order cumulant is analyzed and applied the general band of the high-order cumulant to detect the small targets [16]. Although the detection effect was optimized to some extent, it had a strong dependence on the data set, and its robustness was poor. The singular value decomposition technology was applied to the convolution feature compression processing to reduce the calculation and storage requirements of the model, and the multiscale training method was adopted to adapt to the change of the scale of aviation targets, but there was still a large rate of missed detection, and the detection rate was seriously affected

[17]. The main reasons for missed detection and false detection are that the target object is not only interfered by the luminance, occlusion, and other factors but also affected by the interference factors such as the small scale of the target and the large scale change, the complex and changeable background, and there are many background objects which are very similar to the target.

In this paper, we proposed small target detection algorithm based on migration study and separable network to solve the high rate of false positives, poor robustness, and low detection rate under battlefield environment. The key contributions in our work can be summarized as follows:

(1) In order to strengthen the relationship between the shallow layer and the deep layer, three fusion layers under the idea of the fusion feature pyramid are proposed. The new feature layer obtained by fusion is taken as the input of the next layer to learn the feature extraction. In addition, the deep separable convolutional network is used for feature extraction to reduce the computational load

(2) Adopt exponential linear element (ELU) instead of traditional ReLU activation function. It achieves the effect of BN layer and reduces a lot of computation. At the same time, it is more robust to the noise of input change and has low complexity

(3) An improved Softmax loss function, namely, F-Softmax, is proposed. By introducing angle constraint, the distance between classes can be increased, and the distance within classes can be reduced by strict decision conditions. This will make the classification more accurate. The introduction of key factors can reduce the weight of samples which are easy to classify and meanwhile make the model focus on the samples which are difficult to classify during training. To do this can solve the unbalance problem of positive and negative sample

2. Design of Small Target Detection Architecture

2.1. Network Structure. The Mask R-CNN [18] model is a pyramid-like structure, but the shallow feature map with a large field of view does not have the detailed information of the deep feature map, and the deep feature map of small field view cannot cover the target information. So the Mask R-CNN is not good for small target detection. Meanwhile, the huge calculation of Mask R-CNN model makes the positioning and classification speeds slow. In order to enhance relationship between shallow features and deep features, this paper proposes Mask R-CNN model fused with feature pyramids, as shown in Figure 1.

In this paper, three layers of Conv7_1, Conv8_1, and Conv9_1 are selected for feature pyramid fusion structure. Conv7_1 is fused with Conv6_2, Conv8_1 is fused with Conv7_2, and Conv9_1 is fused with Conv8_2. Next, the fusion calculation of Conv7_1 and Conv6_2 is taken as an example for analysis as shown in Figure 2. The other two fusion methods are the same.

The characteristic in Conv7_1 is fused with the characteristic in Conv6_2. Low level feature Conv6_2 needs to change the number of channels through a 1×1 convolutional layer to reduce the dimension of the feature graph. Similarly, for Conv7_1, the number of channels should be changed through a 1×1 convolutional layer to change it into $19 \times 19 \times 256$. Then, the image size of Conv7_1 should be expanded twice by using bilinear interpolation algorithm to become $38 \times 38 \times 256$. Finally, the low level features and high level features are fused to get a new feature layer. The new feature layer obtained by fusion is taken as the input of the next layer to learn the feature extraction.

2.2. Transfer Learning Structure. This paper introduced the transfer learning combined with CNN to propose a remote sensing image target model recognition algorithm based on transfer learning. Among them, the source domain is PASCAL VOC2012 data set of ten type of targets. Source task focused on image classification in the source domain. The target domain refers to the PASCAL VOC2012 data set of five small type of targets.

Target task is to classify small targets in the image of the land battlefield. The overall structure of the transfer learning method is shown in Figure 3.

The framework of transfer learning used in the algorithm is based on the Mask R-CNN network model, each of which includes multiple convolutionlayers, activation layers, pooling layers, and fully connected layers. The algorithm can be divided into two stages: the preliminary training stage and parameter fine-tuning stage. First, classified training with five small types of targets in VOC2012 data set and get the classification model. Then, under this basis model, classification training is carried out in the other kinds of target image.

The network model used in this paper includes 13 convolution layers, 13 activation function layers, and 4 pooling layers. Among the convolution layer, the convolution kernel size is 3, zero complement is 1, and step size is 1. Among the pooling layer, the window size is 2, and step size is 2. The activation function used in the model is ReLU activation function. The full connection layer improves overfitting by using dropout, which randomly sets the neurons in the model to 0 with a 50% probability to reduce the dependence of fixed connections between neurons. The classification layer adopts NMS function. There are 10 categories of labels, and each is predicted as the probability of the corresponding category.

2.3. Deeply Separable Network. Using the concept of Xception model to balance the accuracy and speed and meanwhile realize extract the attention feature of the image. In the template feature extraction network and feature extraction to be detected network, using depth separable convolution instead of traditional convolution kernels, which means to build the DS-AlexNet (Depthwise Separable-AlexNet) network. And the other module of the network is not need to be changed. To do this can reduce the cost of the network parameters while not affect the accuracy of model.

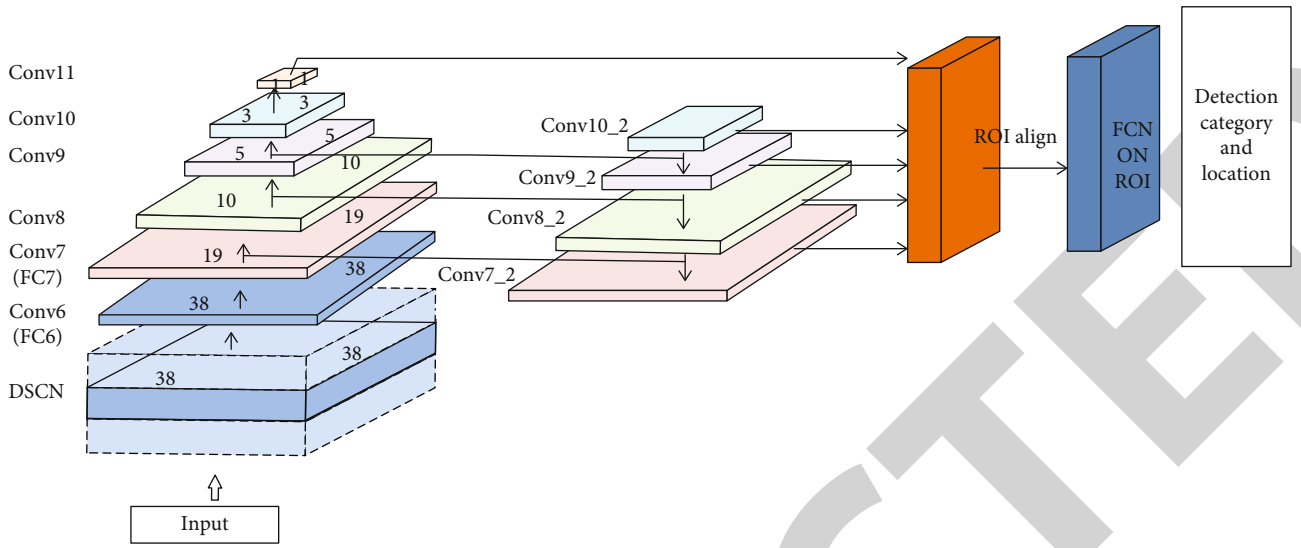


Figure 1: Overall network structure.

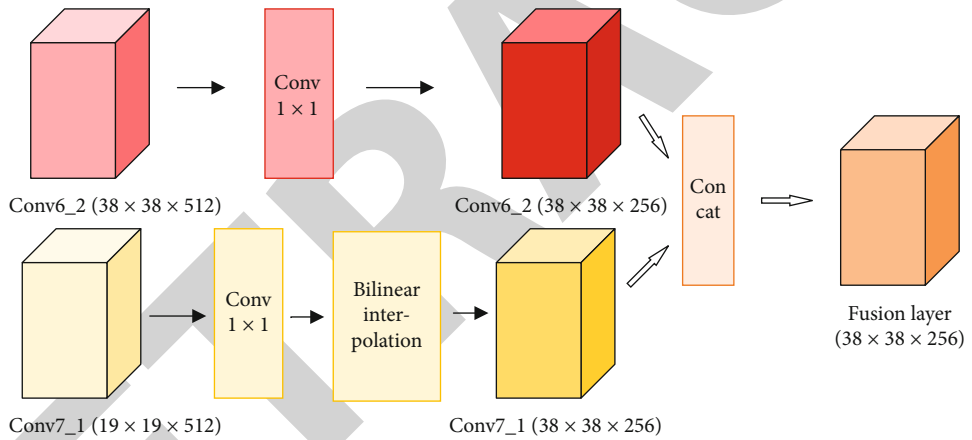


Figure 2: Fusion structure.

The standard convolution structure is shown in Figure 4. Separable convolution is a one-dimensional channel convolution kernel and a two-dimensional position convolution kernel. The channel information and position information in the image are, respectively, learned. The separable convolution structure is shown in Figure 5.

The main purpose of using separable convolution is to separate the spatial cross-correlation information from the channel cross-correlation information, so as to improve the recognition rate while speeding up calculation. Assume the input feature size is $D_k \times D_k \times M$, M is the number of input channels. The standard convolution kernel is $D \times D \times N$, where D is the length and width of the convolution kernel and N is the number of output channels. The calculation amount of a standard convolution is shown in Equation (1).

$$D_k \times D_k \times M \times N \times D \times D. \quad (1)$$

In the case of separable convolution, $D \times D$ filters are applied to M input channels, i.e., $D \times D \times M \times DK \times DK$, and N $1 \times 1 \times M$ convolution filters are applied to combine M input channels into N output channels, i.e., $M \times N \times DK \times DK$. Merge each value in the $1 \times 1 \times M$ feature graph together, and the calculation amount is shown in Equation (2).

$$D \times D \times M \times 1 \times D_k \times D_k + 1 \times 1 \times M \times N \times D_k \times D_k. \quad (2)$$

Compared with the standard convolution structure, such a separable convolution structure requires less computation, as shown in Equation (3).

$$\frac{D \times D \times M \times 1 \times D_k \times D_k + 1 \times 1 \times M \times N \times D_k \times D_k}{D \times D \times M \times N \times D_k \times D_k} = \frac{1}{N} + \frac{1}{D^2}. \quad (3)$$

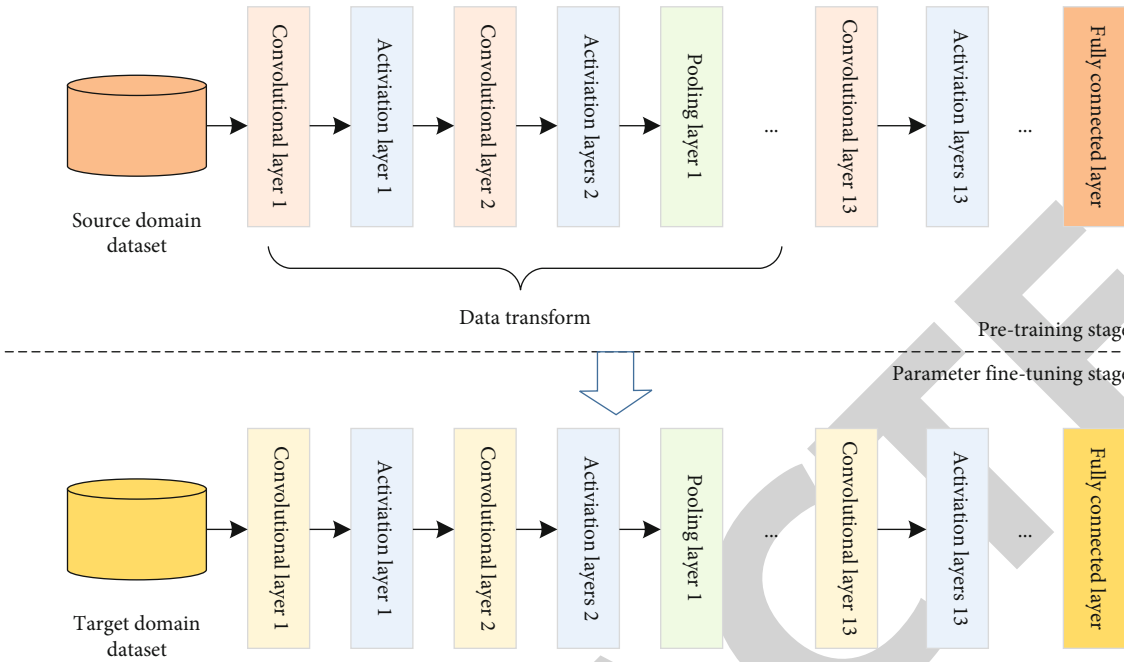


Figure 3: The overall structure of the transfer learning.

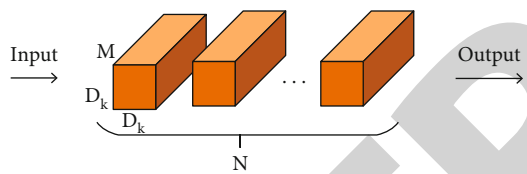


Figure 4: Standard convolution.

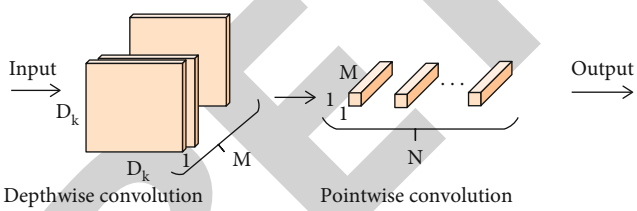


Figure 5: Separable convolution.

Taking Conv4 to Con5 as an example, the feature mapping of the input is $(13 \times 13 \times 384)$, and the standard convolution kernel is $(3 \times 3 \times 256)$.

$$\frac{13 \times 13 \times 384 \times 1 \times 3 \times 3 + 1 \times 1 \times 256 \times 384 \times 3 \times 3}{13 \times 13 \times 384 \times 256 \times 3 \times 3} \approx \frac{1}{128}. \tag{4}$$

As shown in Equation (4), the amount of calculation is reduced to 1/128 of the original amount.

Through the above analysis, it is proved that using the deep separable convolution structure instead of the traditional convolution structure can speed up the calculation and reduce the used computing resources under the condition of ensuring the same feature extraction effect.

2.4. Activation Function and Loss Function

2.4.1. Activation Function. The exponential linear unit (ELU) is used to replace the traditional ReLU activation function, and the ELU function expression is shown in Equation (5).

$$f(x) = \begin{cases} x & \text{if } x \geq 0, \\ \alpha & (e^x - 1), \text{if } x < 0, \end{cases}$$
$$f'(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ f(x) + \alpha, & \text{if } x < 0. \end{cases} \tag{5}$$

The ELU function is an improvement over ReLU. When the parameters are greater than or equal to 0, the computational complexity is low, and the learning speed is fast without the need for exponential operation, which also increases the nonlinear characteristics of the model. When the parameter is less than 0, a smooth function is used instead of the original identity 0, so that the average output value of the activation function is close to zero; therefore, the convergence speed is faster. The BN layer effect is achieved, and a lot of computation is reduced. At the same time, it is more robust to the input noise and has lower complexity.

2.4.2. Loss Function. The classification function used in the Mask R-CNN model is Smooth L1, which approximates the output as a probability distribution, as shown in Equation (6).

$$L_{\text{loc}}(x, l, g) = \sum_{i \in \text{Postive}}^{N} \sum_{m \in \{cx,cy,w,h\}} x_{ij}^{k} \text{smooth}_{L1}\left(l_i^m - \widehat{g}_j^m\right). \quad (6)$$

This paper proposes an improved Softmax loss function with a period of increasing angle constraints and a key factor named F-Softmax.

(1) Angle constraints

The output of the model is $w_1x, w_2x \cdots w_nx$; the classification estimation probability obtained after the loss function is shown in Equation (7).

$$h_w(x) = \frac{1}{e^{w_1x} + e^{w_2x} + \cdots + e^{w_nx}} \begin{bmatrix} e^{w_1x} \\ e^{w_2x} \\ \cdots \\ e^{w_nx} \end{bmatrix}. \quad (7)$$

x in formula (7) represents the sample, n represents the category, and e^{w_nx} represents the parameter weight vector of different sample types. If the input sample x belongs to the n category, then the value of e^{w_nx} is the largest; that is, w_nx is also required to be the largest. Expand the equation to obtain Equation (8).

$$\|w_n\|\|x\| \cos \theta_n > \{\|w_1\|\|x\| \cos \theta_1, \|w_2\|\|x\| \cos \theta_2 \cdots\}. \quad (8)$$

θ is the included angle between the sample probability vector and the parameter weight vector. Assuming an integer n, Equation (9) can be obtained according to the properties of cos function.

$$\|w_n\|\|x\| \cos \theta_n > \{\|w_1\|\|x\| \cos \theta_1, \|w_2\|\|x\| \cos \theta_2 \cdots\} \quad \text{where}\left(\theta_1, \theta_2 \in \left[0, \frac{\pi}{n}\right]\right). \quad (9)$$

The inclusion of the limiting conditions in the formula makes the discrimination more strict, so that in the original loss function, if there is A class of targets that may belong to class A or may belong to class B. At this time, when judging the category of the target, not only the probability vector is required to be the same as the parameter weight vector but also the constraint condition of an included angle is added. Strict criteria can make the distance between classes larger and the distance within classes smaller, so that the classification is more accurate.

(2) Key factors

Angle constraint considers that the distance between the class and class does not take into account the balance of positive and negative samples, on the basis of the previous section introduced the Focal Loss of ideological building loss function, the Focal Loss formula such as type of (10).

$$\text{Loss}_{\text{FL}} = -\alpha_t(1 - p_t)^{\gamma} \log(p_t), \quad (10)$$

where p_t is equal to (11).

$$P_t = \begin{cases} p, \text{if} \quad y = 1, \\ 1 - p, \text{else}. \end{cases} \quad (11)$$

α_t is called the weighting factor; y is called the key factor; $(1 - p_t)$ represents the probability of belonging to the label, within the range of [0,1]; and P is the probability of the target predicted by the model, within the range of [0,1].

Finally, the loss function formula is substituted into Focal Loss, which is the proposed loss function formula F-Softmax, as shown in Equation (12).

$$L_{\text{FS}} = \frac{1}{N} \sum_i -\alpha_t(1 - p_t)^{\gamma} \log(p). \quad (12)$$

The optimal initial value selection is given through experiments, and $\alpha = 0.25$ and $\gamma = 2$ are set. Where n is the number of categories and p is the probability precalculated by Softmax function, the calculation formula is shown in Equation (13).

$$p = \frac{e^{\|x_i\|\|w_i\| \cos \theta_n}}{e^{\|x_i\|\|w_i\| \cos \theta_n} + \sum_{i \neq y_i} e^{\|x_i\|\|w_i\| \cos \theta_n}}, \theta \in \left[0, \frac{\pi}{n}\right]. \quad (13)$$

3. Experimental Results

3.1. Introduction to the Experimental Environment. In the Ubuntu 16.04 operating system, the algorithm in this paper adopts the deep learning framework PyTorch to realize the ground-field target detection algorithm based on the multilevel feature pyramid. The experimental platform uses CPU: Intel(R) Core(TM) I5-8600 3.10 GHz; Memory: 16 G; GPU: NVIDIA GTX 1080TI, training and testing the network in the above environment. In order to verify the accuracy and real-time performance of the algorithm, YOLO v3, Faster R-CNN, and Mask R-CNN algorithms with better current performance were selected for comparison, all of which were tested in the same environment. The training set is made by randomly extracting 70% data from the data set, while the test set is made by randomly extracting 30% data from the data set.

3.2. Introduction to the Experimental Data Set. The target included tank, person, gun, cannon, helicopter, and car. The data set contains 9,000 images of the abovementioned target. Then, the data was expanded to 27,000 by adding noise and scaling to some extent. We also found 3000 relevant video images containing the target from the network. So the data set consists of total 30,000 images. Each image is manually annotated in accordance with the format of PASCAL VOC data set. Some images of the data set are shown in Figure 6.

3.3. Ablation Experiment. The experimental data set is expanded self-made data set. The data set was taken as input, and the parameters of network training were set as follows:

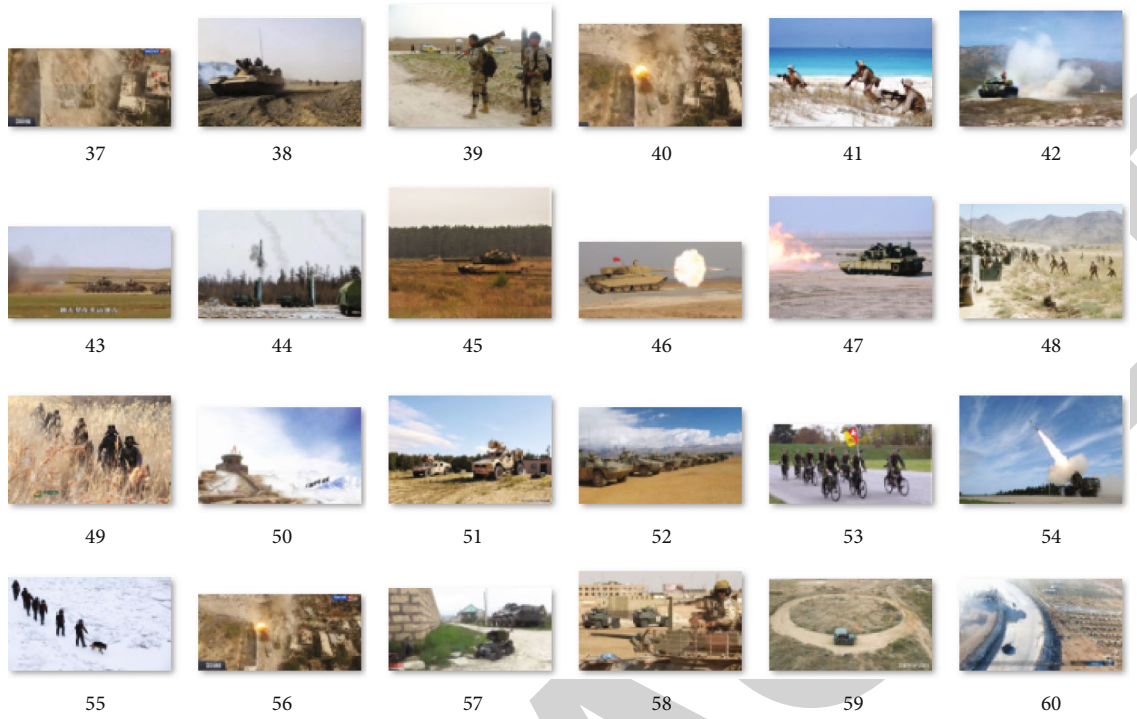


Figure 6: Part of the data set.

learning rate 0.1, the learning rate decreased by 1 order of magnitude after each epoch, regular term $f = 0.1$, and Batch-Size 100. After each epoch, the data set was rearranged randomly.

3.3.1. Activation Function. The recognition accuracy of the network on the test set-the iteration step curve (acc-step) and the loss function value-the iteration step curve (loss-step) are shown in Figures 7(a) and 7(b). It can be seen that with the update of iteration step, the overall recognition accuracy finally reached more than 90%, and the curves of the identification accuracy and the loss function value of the data set basically leveled off after about 3300 iterations.

The activation functions in the residual block are, respectively, set to ReLU and LReLU. The activation functions are shown in Equations (5), (14), and (15). Only the activation function in the model is changed; other parts of the model remain unchanged and are trained under the same training set.

a is the adjustment parameter, and it control the activation of the ELU function in the negative half axis.

$$\text{Re LU}(x) = \begin{cases} x, & \text{if } x > 0, \\ 0, & f\ x \le 0, \end{cases} \tag{14}$$

$$\text{L Re Lu}(x) = \begin{cases} x_i, & \text{if } x_i \ge 0, \\ \alpha_i x_i, & \text{if } x_i < 0. \end{cases} \tag{15}$$

a_i is fixed. i means different channels correspond to different a_i.

It can be seen from Figure 8 that the convergence of LReLU is close to ELU during 20000-25000 iterations. However, compared with ReLU and LReLU, as the number of iterations increases, ELU has the minimum final loss function value, and the training effect of the model is better.

3.3.2. Comparison of Loss Functions. In order to verify the superiority of F-Softmax function, ROC curve is used to evaluate the influence of various loss functions on the classification of model samples. Softmax loss function and cross-entropy loss function are used to compare with F-Softmax function. In order to ensure objectivity, other parameters of the model remain unchanged.

From Figure 9, the classification effect of the Softmax is the worst. The cross-entropy loss function can effectively solves the probability problem of multiple classifications and improves the classifier effect. The F-Softmax function not only effectively solves the guidance problem of difficult samples and simple samples but also effectively deals with the problem of sample imbalance, making the loss function more reasonable. The classification model using the loss function has the best classification performance.

3.3.3. Comparative Experiment of Transfer Learning. In order to verify the applicability of the migration study, this paper is the first on the five types of self-made data set to train and get classification model. Then, using this model to all the 10 kinds of target image data set to train the model, the model will give recognition rate and loss function change curve of the two training condition, respectively. The five types of data sets used to train the original model include person, armoured vehicles, gun, tank, and drone. The other

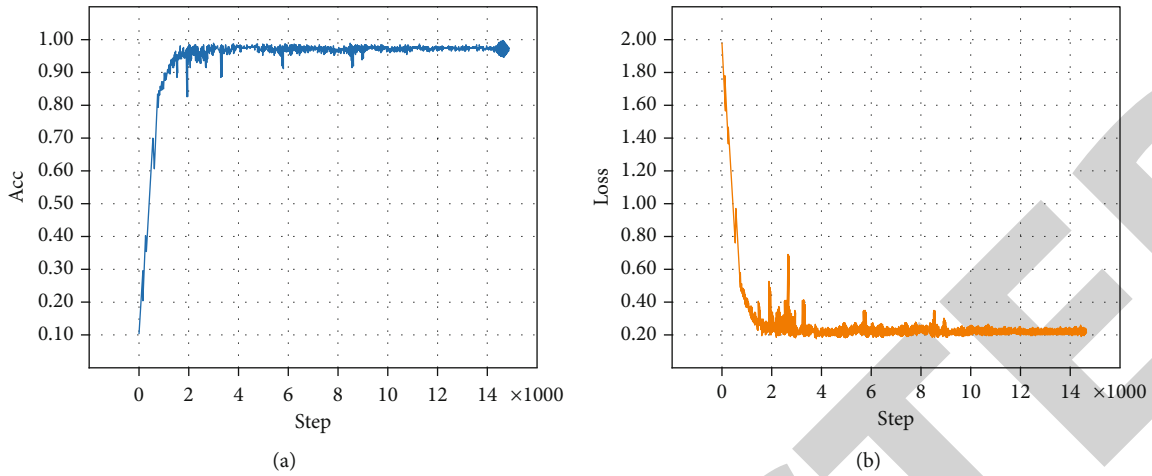


Figure 7: (a) Acc-step curve. (b) Loss-step curve.

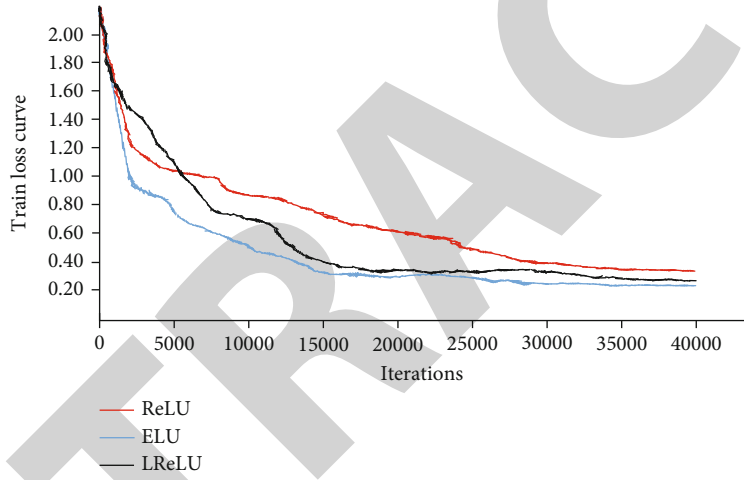


Figure 8: Loss function changes.

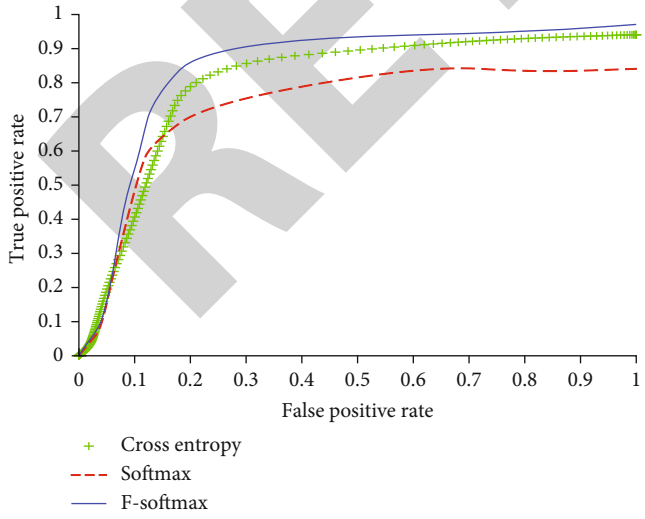
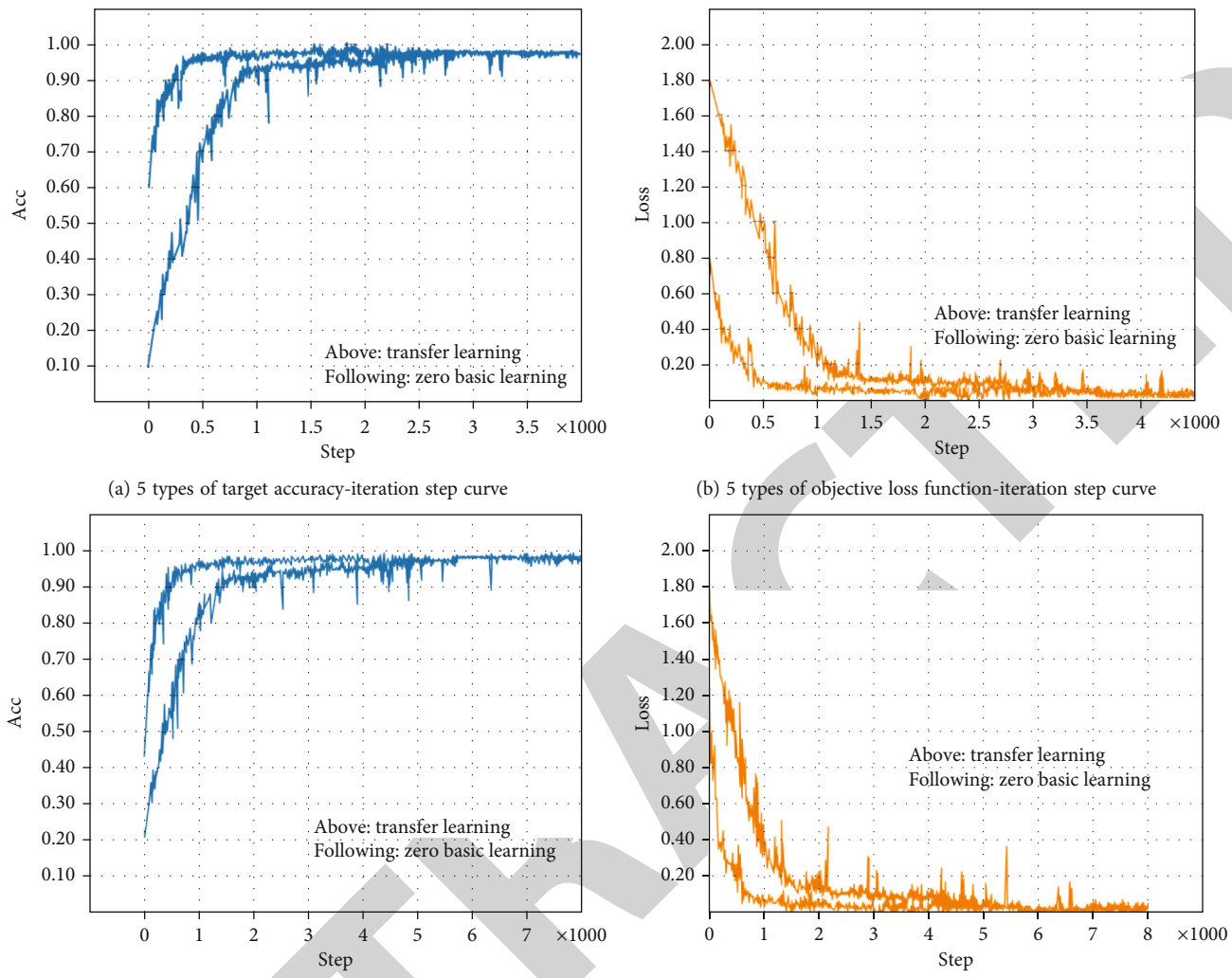


Figure 9: ROC curve.

five data sets used to verify transfer learning include knives, helicopter, car, bulldozer, and cannon, and the other 10 data sets include all of the above targets. Figure 10 shows the comparison of classification accuracy and loss function curves of the model under zero-based learning and transfer learning modes.

The parameters of the network model were set as follows: the learning rate was 0.1, the regular term $f = 0.1$, and the BatchSize was 100. After each epoch, the data sets were rearranged randomly, and the amplified self-made data set was used as input to train the network. As can be seen from Figures 10(a) and 10(b), when the five types of data sets are classified, the initial value of classification accuracy of zero-based learning is 11.6% while the transfer learning method in the same period is as high as 62.3%. After about 1800 steps, the accuracy of network classification in the transfer learning mode reached a peak of more than 90%, and after about 4500 steps, the accuracy curve had no obvious change. The classification accuracy and loss function curves of the model combining the transfer learning on all 10 types of targets are shown in Figures 10(c) and 10(d).



(a) 5 types of target accuracy-iteration step curve

(b) 5 types of objective loss function-iteration step curve

(c) 10 types of target correct rate-iteration step curve

(d) 10 types of objective loss function-iteration step curve

Figure 10: Comparison of experimental results of transfer learning and zero-based learning.

The initial accuracy values are 21.3% and 44.1%, respectively. After about 2800 steps, the classification accuracy in the transfer learning mode was higher than 92%, slightly higher than the 90% in the zero-basic learning mode of 4800 steps. The iteration curves of transfer learning mode in the above two training sets are smoother, and the model training speed is faster.

3.3.4. Model Comparison Experiment. In this section, the existing YOLO-v3, Faster R-CNN [19], and Mask R-CNN methods will be used to compare with our proposed algorithm which called Ours+. Ours+ is a model for transferring 5 types of data training. All methods are trained in the same data set and tested in the same test set. There are eight parameters used to compare the performance of the four algorithms: mAP (mean Average Precision), AP1 (Average Precision of tank), AP2 (Average Precision of person), AP3 (Average Precision of gun), AP4 (Average Precision of cannon), AP5 (Average Precision of helicopter), AP6 (Average Precision of car), and FPS (Frame Per Second).

Table 1: Test data.

	Ours+	YOLO v3	Faster-RCNN	Mask R-CNN
mAP	66.5	59.4	65.7	66.0
AP1	68.3	61.2	66.7	68.0
AP2	65.5	57.3	65.3	65.1
AP3	65.3	56.5	65.0	64.5
AP4	66.2	59.0	64.8	65.0
AP5	67.0	61.5	66.3	66.9
AP6	66.7	60.9	66.1	66.5
FPS	14	35	15	12

As shown in Table 1, compared with YOLO v3 model, as the method mainly focuses on lightweight detection so it is the fastest among the four models compared in terms of FPS, but the detection accuracy is far behind the other methods. The original design intention of Faster R-CNN and Mask R-CNN is two-stage structure, which have candidate

Figure 11: Part of the test results.

region generation network. So the network is far more complex than other methods, and the detection accuracy is super than YOLO v3. The proposed method improves several shortcomings of Mask R-CNN model, so the accuracy of the FPS is super than Mask R-CNN. Some test results are shown in Figure 11.

4. Conclusion

To solve the problem of low accuracy of small target detection, this paper proposes a small target detection algorithm based on transfer learning and deep separable network. Firstly, feature extraction is carried out by deep separable convolutional network, which reduces the amount of computation. Then, the feature pyramid fusion structure is used to fuse the high-level and low-level feature information, optimize the shallow feature information of the network, and effectively compensate for the loss of information caused by continuous pooling, so as to extract more shallow detail texture information and improve the detection performance of small targets. Finally, the activation function and loss function are optimized to solve the imbalance of positive and negative samples, so as to optimize the network performance. The whole network model is trained by transferring the learning method, and experiments are carried out on the PASCAL VOC2012 data set. The experimental results show that the proposed model is significantly better than other algorithm models in the detection accuracy of small targets.

Data Availability

The processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Conflicts of Interest

These are no potential competing interests in our paper.

Authors' Contributions

All authors have seen the manuscript and approved to submit to your journal.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62171360), the Science and Technology Program of Xi'an Technology Department (2020KJRC0037), and the Principal Research Foundation of Xi'an Technological University (XGPY200217).

References

[1] X. Bai and F. Zhou, "Analysis of new top-hat transformation and the application for infrared dim small target detection," *Pattern Recognition*, vol. 43, no. 6, pp. 2145–2156, 2010.

[2] P. Wang, M. Sun, H. Wang, X. Li, and Y. Yang, "Convolution operators for visual tracking based on spatial-temporal

regularization," *Netural Computing and Applications.*, vol. 32, no. 10, pp. 5339–5351, 2020.

[3] C. Gao, "Infrared small-dim target detection based on Markov random field guided noise modeling," *Pattern Recognition the Journal of the Pattern Recognition Society*, vol. 76, pp. 463–475, 2018.

[4] P. Wang, J. Wu, H. Wang, and X. Li, "Low-light-level image enhancement algorithm based on integrated networks," *Multimedia Systems*, vol. 7, 2020.

[5] X. Bai and Y. Bi, "Derivative entropy-based contrast measure for infrared small-target detection," *IEEE Transactions on Geoence and Remote Sensing*, vol. 56, no. 4, pp. 2452–2466, 2018.

[6] P. Głomb, K. Domino, M. Romaszewski, and M. Cholewa, *Band Selection with Higher Order Multivariate Cumulants for Small Target Detection in Hyperspectral Images*, 2018, https://arxiv.org/abs/1808.03513.

[7] S. Chen, F. Luo, C. Hu, and X. Nie, "Small target detection in sea clutter background based on Tsallis entropy of Doppler spectrum," *Journal of Radars*, vol. 8, no. 3, pp. 344–354, 2019.

[8] L. Hua, S. Yu-Long, and Q. Feng, *Detection of Small Target in Aerial Photography Based on Deep Learning*, Chinese Journal of Liquid Crystals and Displays, 2018.

[9] W. Sun, D. Yan, J. Huang, and C. Sun, "Small-scale moving target detection in aerial image by deep inverse reinforcement learning," *Soft Computing*, vol. 24, no. 8, pp. 5897–5908, 2020.

[10] L. Liangkui, W. Shaoyou, and T. Zhongxing, "Using deep learning to detect small targets in infrared oversampling images," *Journal of Systems Engineering & Electronics*, vol. 29, no. 5, pp. 947–952, 2018.

[11] Y. Li, W. Wang, and Y. Jiang, "Through wall human detection under small samples based on deep learning algorithm," *IEEE Access*, vol. 6, pp. 65837–65844, 2018.

[12] A. Mansour, W. M. Hussein, and E. Said, "Small objects detection in satellite images using deep learning," in *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, IEEE, 2020.

[13] L. I. Dajun, H. E. Weilong, and G. Bingxuan, *Building Target Detection Algorithm Based on Mask-RCNN*, ence of Surveying and Mapping, 2019.

[14] P. Wang, F. Huitong, X. Li, J. Guo, Z. Lv, and R. Di, "Multi-feature fusion tracking algorithm based on generative compression network," *Future Generation Computer Systems*, vol. 124, pp. 206–214, 2021.

[15] A. O. Vuola, S. U. Akram, and J. Kannala, "Mask-RCNN and U-net ensembled for nuclei segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, 2019.

[16] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: new benchmark and state of the art analysis," *In CVPR*, vol. 8, 2014.

[17] T. Y. Lin, P. Dollár, and R. Girshick, "Feature pyramid networks for object detection," *IEEE Computer Vision and Pattern Recognition*, vol. 11, 2017.

[18] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Conference on Computer Vision*, vol. 12, 2017.

[19] 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.