Infrared small target detection method based on multi-scale feature fusion

Fang Wang¹,², Chuanqiang Li³, Bo Wu¹,³*, Kun Yu ³ and Chan Jin²

¹ College of Electronic and Electrical Engineering, Henan Normal University, Xinxiang 453007, China
² Key laboratory of interfacial Physics Technology project, Chinese Academy of Sciences, Shanghai 201800, China
³ Henan Key Laboratory of Optoelectronic Sensing Integrated Application, Xinxiang 453007, China

*Corresponding author’s e-mail: wubo@htu.edu.cn

Abstract: Infrared small target detection is widely used in aerial target detection and tracking system because of its long detection distance and strong anti-jamming ability. Aiming at the shortcomings of the current infrared small target detection algorithm, such as low precision rate and high false alarm rate when dealing with complex background. In this paper, we propose an end-to-end infrared small target detection model (called MFSSD) based on multi-scale feature fusion. Considering the trait of the targets, we propose a feature fusion module by using a refinement and fusion feature map method, and improve the correlation of different channels through the SP Module. The experimental results of three different sequences of infrared image detection show that the average detection accuracy of the MFSSD algorithm in infrared small target detection is as high as 87.8%. Compared with the traditional multi-scale feature fusion detection algorithm, both the precision rate and the recall rate have been significantly improved.

1. Introduction
Infrared small target detection is the key technology of aerial target detection and tracking system, and is widely applied in infrared search and tracking system [1], infrared early warning [2] and infrared monitoring. Early prediction of the location and distance of small targets is crucial. However, the infrared image is greatly affected by changes in the surrounding environment, Meanwhile, the target imaging distance in the infrared image is relatively long and appears in the form of dots. It cannot express various information such as the texture and shape of the target well. It is more difficult to accurately locate the position of the target. Therefore, infrared small target detection is favored by the majorities of researchers [3].

When it comes to CNNs, it must be AlexNet, which achieved great success in the ImageNet dataset classification task in 2012. In the following years, there was a rapid upsurge of CNNs. The use of CNNs for target detection became the mainstream, and then some classic network structures appeared, such as VGG [4], ResNet and GoogLeNet [5].

From the current development of CNNs, target detection was mainly divided into two categories: one-stage target detection algorithms and two-stage target detection algorithms. Among them, the classical two-stage target detection algorithms mainly include: R-CNN[6], Fast-RCNN[7] and Faster-RCNN[8]. These algorithms first extract candidate frames using traditional selective search methods,
fix the candidate frames to a uniform size and input them into CNN for feature extraction, and then use classifiers for classification and positioning. Although this method improves the detection ability, it also has some limitations, due to the separation of regional selection and feature extraction during the training network, which leads to complex training.

For the shortcomings of the two-stage target detection algorithms, one-stage target detection algorithms has been proposed, such as YOLO[9], SSD[10]. One-stage target detection algorithms mainly takes advantage of the characteristics of the convolutional neural network, which can directly input an image, extract features through convolution and pooling, and then classify and position them by the regression method, which greatly improves the detection speed. Additionally, the detection ability of small targets is insufficient. The SSD algorithm combines Faster-RCNN and YOLO, using multi-scale feature maps to predict the classification and positioning of targets at the same time, which is more suitable for the detection of targets of different sizes. Because of the advantages of SSD, researchers are more inclined to combine the SSD algorithm with infrared small targets to realize small target detection.

2. Methodology

In this paper, we design an MFSSD algorithm, which was a multi-scale feature fusion infrared small target detection method. As was shown in figure1. The model structure has the trait of high precision and high speed. We firstly introduced the subpixel convolution layer and the pathway layer, and used sub-pixel convolution layer and pathway layer instead of down-sampling and up-sampling. Then the feature fusion module and SP Module were introduced in detail. Finally, a multi-scale infrared small target detection method was formed by combining it with the SSD algorithm.

2.1. Feature map adjustment method

Inspired [11, 12] by these papers, we used the pathway layer and sub-pixel convolution layer to replace the up-sampling and down-sampling. Compared with the previous interpolation and pooling methods, this method mainly completed the corresponding process by rearranging the pixels. The sub-pixel convolution layer was an input of a low-resolution feature maps. Through pixel recombination between multiple channels of the convolution kernel, a high-resolution feature maps can be obtained. It can transform the low-resolution feature graphs $N \times (C \times r \times r) \times W \times H$ into a high-resolution feature graphs $N \times C \times (H \times r) \times (W \times r)$. In the pathway layer, two adjacent pixels are placed in different channels to realize the down-sampling operation. It can transform the high-resolution feature graphs $N \times C \times (r \times H) \times (r \times W)$ into to a low-resolution feature graphs $N \times (C \times r \times r) \times H \times W$, where $N$, $C$, $W$, $H$ and $r$ respectively represent the number, channels, width and height of images as well as the multiple of up-sampling or down-sampling. This method was shown in Figure 2.
2.2 FFM Module
Multiple feature maps were adopted in the SSD algorithm for target detection, and each feature map contained unique feature information. The high-resolution feature map showed the features of large targets clearly, which was suitable for target positioning. Low-resolution feature maps contained more texture information and were suitable for target classification. Therefore, as was shown in Figure3, we introduced the structure of the feature fusion module in detail, S1, S2 and S3 were the original feature graphs directly output by the backbone network with low to high resolution, and P2 was the feature map after feature fusion. To improve the detection ability of different size feature maps to infrared small target. First, S1 used the sub-pixel convolution layer to obtain S1_up through pixel rearrangement, then S3 used the pathway layer to obtain S3_down through pixel rearrangement, where S1_up, S3_down and S2 have the same size, and the channels of S1_up, S3_down and S2 were spliced to get feature map C2. Then, using 1*1 convolution kernel, the amount of channels in the feature map C2 was reduced to the same group of channels as the feature map S2 to obtain the P2 feature graph for target classification and positioning. This process can effectively enrich more semantic and texture information in the fused feature map. In the whole fusion process, the sub-pixel convolutional layer and the pathway layer can automatically learn the mapping relationship through gradient update. Then the convolutional layer was used to adjust the channels of the feature maps used for detection, keeping the same number of channels as the feature maps before fusion. Therefore, this feature fusion module had stronger advantages in detecting infrared small targets.

2.3. SP Module
We added the SP module to each feature map that needed to be targeted for classification and positioning after feature fusion. When the target was located through the feature map after feature fusion, it does not consider the feature information of each channel. In SP Module, the input size of this module was W*H*C. the input feature map was first subjected to average pooling and compressed into 1*1*C. Then, the model can be passed through two connection layers (First reducing the dimensionality through the 1*1*(C/16) convolution kernel, and then conversely increasing the dimensionality through the 1*1*C convolution kernel.). A weight was generated for each feature channel to improve the correlation between each channel of the feature map. Then, to maintain the same size as the input feature map, the feature map \( Y_1 \) was obtained by up-sampling operation. Next, \( Y_1 \) was added to \( Y_0 \) to improve the
useful feature information in the original feature map. Finally, the final output result was used to classification and positioning the target.

![Figure 4. SP Module structure.](image)

3. Dataset and experiment setup

3.1 Image dataset
In this experiment, three sequences of infrared images are adopted, among which the image is 256*256 and the image shooting frequency of each sequence is 100Hz. The image information of the dataset is summarized in Tab.1. Infrared images with a single background, multiple targets and complex backgrounds are included. A total of 1497 images were selected. A total of 1297 images were randomly selected as the training set, and another 200 images were selected as the test set.

| Name   | Total number | Image resolution | Detail                                           |
|--------|--------------|------------------|--------------------------------------------------|
| Data1  | 399          | 256×256          | The background is a sky background with varying degrees of thermal noise and a single target |
| Data2  | 100          | 256×256          | Background is the intersection of sky and ground background, a single target               |
| Data3  | 998          | 256×256          | The background is a sky background with two targets and cross flying                        |

3.2 Hardware configuration
We experimented with an Intel Xeon E3-1220 V6 @3.00Ghz quad-core processor with 32G of RAM and a Nvidia GTX2080 Ti. All experiments were carried out with CUDA 10.2 on PyTorch1.5.1, Python 3.7, and Window 7 platforms. The MFSSD algorithm was written by PyTorch framework in Windows 7 environment.

3.3 Training parameters
The MFSSD algorithm is trained by setting the following parameters. Our input image is uniformly fixed at 256*256, and we will use the above infrared small target dataset for training and testing. We set the batch to 32. We initially set the learning rate to 0.0005 and reduce the learning rate by 10 times per iteration 50 times, and the number of iterations to 200. Then, we trained and tested the original SSD algorithm, trained and tested it in a progressive way, and finally obtained the results of our MFSSD algorithm. All the above algorithms have the same training parameters.

4. Detection result
In this study, we used a dataset of infrared small targets to evaluate the algorithm. By comparison with Tab.2 algorithm, the advantages of the MFSSD algorithm proposed were reflected.
Table 2. The comparison algorithm needed in the experiment.

| Model | Model description |
|-------|-------------------|
| 1     | SSD               |
| 2     | SSD+FFM (FFM Module adopts up-sampling and down-sampling methods for fusion) |
| 3     | SSD +FFM (FFM Module adopts subpixel convolutional layer and path layer methods for fusion) |
| 4     | SSD + FFM (FFM Module adopts up-sampling and down-sampling methods for fusion) + SP Module |
| 5(our)| SSD+FFM (FFM Module adopts subpixel convolutional layer and path layer methods for fusion) + SP Module |

4.1 Testing result

At present, the end-to-end target detection algorithms mainly included YOLO, SSD and other algorithms. The detection of targets was greatly improved both time and accuracy. The MFSSD algorithm proposed in this paper mainly used the multi-scale detection idea of the SSD algorithm for detection. Therefore, we use the row-order asymptotic method to compare the Model 1. Table 3 showed the parameter comparison results of testing with different model structures. It was found that Model 5 Map can reach 87.5%. Compared with Model 1, Model 5 had obvious advantages.

Table 3. Test results of different networks.

| Model | Input | Train | Test | Map  | Fps |
|-------|-------|-------|------|------|-----|
| 1     | 256   | 1297  | 200  | 82.5 | 29  |
| 2     | 256   | 1297  | 200  | 85.5 | 23  |
| 3     | 256   | 1297  | 200  | 86.2 | 25  |
| 4     | 256   | 1297  | 200  | 86.1 | 15  |
| 5     | 256   | 1297  | 200  | 87.8 | 17  |

To compare the effect of different methods to realize the up-sampling and down-sampling on the detection accuracy of the model, we first used the interpolation method and maximum pooling method [31] to construct the feature fusion module. We found that the detection accuracy was improved by 3% compared with that of the Model 1. Then, when we adjusted the size of the feature maps in the feature fusion module, we used the sub-pixel convolution layer and the path layer. We found that the detection accuracy was improved by 3.6% compared with that of the Model 1. Compared with Model 2 and Model 3, we found that the detection accuracy of sub-pixel convolutional layer and path layer was improved by 0.7% instead of the interpolation method and maximum average pooling method.

By comparing Model 3 and Model 5, we can clearly see the effectiveness of the SP module. The SP Module can automatically increase the expression ability of useful information and suppress the expression ability of useless information by adding a small amount of computation. Adding the SP Module to the existing model structure increases the MAP by 1.6%. This finding suggested that the SP Module was useful for learning the characteristic information between different channels.

Finally, the Figure 5 was a line diagram comparing the precision and recall of the Model 1 with the Model 5. As was shown in the Figure 5, when recall rate was below 0.2, the difference between the two algorithms was not significant. However, when recall ratio is between 0.2 and 1, the Model 5 had obvious improvement in both precision and recall.
5. Conclusion

In this paper, an infrared small target end-to-end detection model (called MFSSD) based on multi-scale feature fusion was proposed. The main network of this algorithm was derived from the classification network. Because the detection targets were infrared small targets, and the target detection was different from the target classification. We first integrated shallow and deep texture information into the feature map that needed to be used for detection, and improved the accuracy of small target detection. Next, the sub-pixel convolution layer and the pathway layer instead of the network for up-sampling and down-sampling were used to resize the size of the feature map. Finally, to learn the feature information of different channels more effectively, the SP Module was added to the feature map after fusion, it can assign weight of corresponding size to each channel of feature graph with different feature information. The results of the experiment found that the MFSSD algorithm in infrared small target detection was as high as 87.8%. In addition, the MFSSD algorithm had a strong adaptability to targets with different backgrounds, and the MFSSD algorithm was far superior to the SSD algorithm in terms of precision and recall, which makes it of great research value in the military field. In this study, due to the increase of algorithm complexity, the recognition speed was reduced. The next work was to improve the recognition speed under the condition that the accuracy remains unchanged.

References

[1] X. Shao, H. Fan, G. Lu, and J. Xu, (2012)"An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system," Infr. Phys. Technol. 55, 403-408.

[2] P. Coppo, (2015) "Simulation of fire detection by infrared imagers from geostationary satellites," Remote Sens. Environ. 162, 84-98.

[3] Y. Bi, X. Bai, T. Jin, and S. Guo, (2017)"Multiple Feature Analysis for Infrared Small Target Detection," IEEE Geosci. Remote Sens. Lett. 14(8), 1333-1337.

[4] K. Simonyan, A. Zisserman, (2014) "Very deep convolutional networks for large-scale image recognition," arXiv Preprint, arXiv:1409.1556.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, (2015) " Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.1–9.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, (2014) "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 580-587.

[7] R. Girshick, (2015)"Fast R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1440-1448.
[8] S. Ren, K. He, R. Girshick, and J. Sun, (2017) "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137-1149.
[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, (2016) "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 779-788.
[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, (2016) "SSD: Single Shot MultiBox Detector," in Proc. Eur. Conf. Comput. Vis. (ECCV) 9905, 21-37.
[11] J. Redmon and A. Farhadi, (2017) "YOLO9000: Better, Faster, Stronger," in Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit., pp.6517-6525.
[12] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, (2016) "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1874-1883.