

Article

Spatial Small Target Detection Method Based on Multi-Scale Feature Fusion Pyramid

Xiaojuan Wang ^{1,2}, Yuepeng Liu ³, Haitao Xu ^{1,2,*} and Changbin Xue ^{1,*} 

¹ National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China; wangxiaojuan21@mailsucas.ac.cn

² University of Chinese Academy of Sciences, Beijing 100049, China

³ School of Mechanical and Electrical, Beijing Institute of Technology, Beijing 100081, China; 3220210129@bit.edu.cn

* Correspondence: xuhaitao@nssc.ac.cn (H.X.); xuechangbin@nssc.ac.cn (C.X.)

Abstract: Small target detection has become an important part of space exploration missions. The existence of weak illumination and interference from the background of star charts in deep and distant space has brought great challenges to space target detection. In addition, the distance of space targets is usually far, so most of them are small targets in the image, and the detection of small targets is also very difficult. To solve the above problems, we propose a multi-scale feature fusion pyramid network. First, we propose the CST module of a CNN fused with Swin Transformer as the feature extraction module of the feature pyramid network to enhance the extraction of target features. Then, we improve the SE attention mechanism and construct the CSE module to find the attention region in the dense star map background. Finally, we introduce improved spatial pyramid pooling to fuse more features to increase the sensory field to obtain multi-scale object information and improve detection performance for small targets. We provide two versions and conducted a detailed ablation study to empirically validate the effectiveness and efficiency of the design of each component in our network architecture. The experimental results show that our network improved in performance compared to the existing feature pyramid.

Keywords: small target detection; attention mechanism; feature fusion pyramid



Citation: Wang, X.; Liu, Y.; Xu, H.; Xue, C. Spatial Small Target Detection Method Based on Multi-Scale Feature Fusion Pyramid. *Appl. Sci.* **2024**, *14*, 5673. <https://doi.org/10.3390/app14135673>

Academic Editor: Antonio Fernández-Caballero

Received: 28 May 2024

Revised: 19 June 2024

Accepted: 21 June 2024

Published: 28 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the continuous development of space technology and the increasing demand of human beings for space exploration, space detection has gradually become a research hotspot in the field of spaceflight [1]. Within this topic, space target detection is the first part of space exploration, through which space targets are detected and recognized to obtain the types and orbital positions of space targets, providing rich information for subsequent orbiting and mission planning [2]. In space target detection missions, targets with large size and signal-to-noise ratios are easily detected. However, for small space targets, such as asteroids [3–5] and space debris [6,7], the presence of a star map background in deep space and poor lighting conditions make the imaging area in the image smaller and weaker in energy, which brings difficulties to target detection.

In recent years, spatial small target detection has gained broad attention and there are a large number of related research results. In a single-frame image, spatially small targets mostly appear as low-intensity light spots and are imaged in a small area with only a few pixels in the image. Feature-based methods such as scale-invariant feature transform (SIFT) [8], histograms of method gradients (HOG) [9], and the local contrast method [10] can enhance the target information while suppressing the interference of background clutter, but for the task of detecting small targets in space, the existing methods' detection capabilities are weak. To address this problem, researchers have proposed a detection method based on sequential images, where the information about the target in the image

is obtained by processing multiple frames. Zhenwei Li et al. [11] used the matched frame difference method to detect small spatial targets and obtained small targets in starry sky images by differentiating the sequential frames and then using a mathematical morphology filtering algorithm to eliminate irrelevant data. K. Fujita et al. [12] proposed an improved optical flow algorithm to estimate an image sequence. J. Xi et al. [13] proposed a multi-stage quasi-hypothesis testing (TMQHT) method to achieve spatial debris detection under low signal-to-noise ratio conditions through time-indexed filtering and multi-stage time indexing. However, all of the traditional methods face the following challenges: firstly, the robustness of traditional methods for target detection under illumination and background interference is poor, and secondly, the existing detection methods need to be executed on each pixel of each image in a sequence of images, and thus, the computational cost is high.

With the rise of artificial intelligence, deep learning has been successful in several fields [14–16], among which computer vision has become a key area of research, and more and more studies are now focusing on spatial target detection. Hu et al. [17] proposed a spatial dark and weak small target detection and tracking method using bilateral filtering and a long short-term memory network. González R. E. et al. [18] utilized the YOLO network trained on self-constructed astronomical datasets for galaxy detection and classification. Jia P. et al. [19] proposed an improved Resnet-50 and feature pyramid network under the Faster R-CNN framework for spatial target detection and classification. Guo et al. [20] proposed a channel and spatial attention CSAU-Net network for real-time dark space target detection and segmentation based on spatial image features.

However, the existing methods do not consider the effects of different illumination and image signal-to-noise ratio conditions on the detection of small targets in space, and this also poses a great challenge to the detection of small targets in space due to the similarity of their intensities with the energies of the background stars.

Aiming at the problems of existing methods, this paper proposes a spatial small target detection method based on a multi-scale feature fusion pyramid to improve the detection performance of spatial small targets. First, we propose the CST (a CNN fused with a Swin Transformer) module of a CNN (Convolutional Neural Network) fused with a Swin Transformer as the feature extraction module of the feature pyramid network to enhance the extraction capability of target features. Then, we improve the SE (Squeeze-and-Excitation) attention mechanism and construct the CSE (Convolution-based Squeeze-and-Excitation Attention) module to find the attention region in the dense star map background. Finally, we introduce the ISPP (improved spatial pyramid pooling) module to fuse more features to increase the sensory field to obtain multi-scale object information and improve detection performance for small targets.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 describes the construction of the dataset and the improvement of the algorithm. Section 4 provides the details of the experiments and test results, which are analyzed and discussed. Finally, Section 5 draws conclusions.

2. Related Work

2.1. Data Augmentation

Data augmentation is a commonly used method in computer vision to expand a dataset, which can improve the robustness of the model to images in different environments. Color transformations are usually achieved by changing the brightness, contrast, and saturation of an image, and for geometric transformations, datasets are usually expanded using operations such as image flipping, panning, and rotation [21]. In addition to the commonly used individual image data enhancement methods, researchers have proposed methods to use multiple images together for data enhancement such as MixUp, CutMix, and AugMix [22]. MixUp mixes any two random images based on the mixing factor and mixes the labels in the same way. AugMix augments the images with a randomized augmentation magnitude, which is then merged to generate a new image. However, the existing data enhancement methods are less effective in small target detection tasks; to

solve this problem researchers have proposed new data enhancement methods such as oversampling, copy-pasting, and mosaic data enhancement.

2.2. Vision Transformers

Currently, the Transformer model is heavily used in natural language processing with success; researchers link CV (Computer Vision) and NLP (Natural Language Processing) domain knowledge to apply the Transformer model to computer vision tasks. CNNs lead to loss of spatial information in convolution and pooling; in contrast, Transformer introduces location information into the model through positional encoding, which introduces location information into the model. ViT [23] divides the image into 16×16 patches and projects each patch as a fixed-length vector to use a Transformer for type prediction, achieving excellent results and providing a new paradigm for visual feature learning. Subsequently, DeiT [24] and PVT [25] improve on ViT to reduce the computational complexity and memory occupation of the algorithm and improve its performance in various downstream tasks. In addition, Swin Transformer is able to capture long-range dependencies between different regions in an image through multi-level windowed self-attention operations, which effectively enhances the modeling capability of the model. Meanwhile, the introduction of windowed self-attention operation can effectively model the spatial features in the image and improve the accuracy of image classification.

2.3. Attention Mechanism

The attention mechanism improves the performance of the model to some extent by assigning different attention weights to different parts of the model inputs and enhancing the learning of key information according to the attention weight distribution model. The existing attention mechanisms mainly include the channel attention mechanism, spatial attention mechanism, hybrid attention mechanism, and self-attention mechanism. The channel attention mechanism improves the expressive ability of the network by constructing the interdependence between channels, such as SENet [26], ECANet [27], and so on. The spatial attention mechanism mainly focuses on different regions of the image and combines the regional information to obtain the image information at different locations such as RAM, STN, etc. The hybrid attention mechanism mixes channel information and spatial information and extracts information features such as CBAM [28], SANet [29], etc. Self-attention mechanisms compute attention on the input sequence itself by mimicking biological visual saliency detection and selective attention to achieve global attention to the input sequence, such as single-head self-attention, multi-head self-attention, and sliding window multi-head self-attention.

3. Method

3.1. Dataset Construction and Processing

In order to ensure the performance of the small target detection algorithm, the space small target database is needed for training, and also, the database is the basis for algorithm performance testing. Existing spatial datasets include the SPEED dataset, URSO dataset SPEED+ dataset, etc. However, the existing aerospace datasets mainly focus on the relative attitude estimation of spacecraft, which does not satisfy our data requirements for small target detection tasks in space. More work has been performed on spacecraft dataset construction in recent years, such as [30,31], among others, who have constructed spacecraft datasets, and no more research has been performed on small targets in space.

To meet our requirements for the database of space small target detection tasks, we constructed data from 8000 space small target images. These include three types of objects, small objects, spacecraft, and space debris, in order to cover as many different types of common natural or man-made small targets in space as possible. The image data were obtained by building a space environment simulation scene in 3D modeling software and rendering it. Firstly, we collected high-precision 3D models of 300 small objects and 10 spacecraft in open sources and constructed an 8000×2000 star map background based

on real space background data. Among them, the former is used to simulate the small targets of small objects and spacecraft in space. Since there are fewer open data on space debris, we edited and modified the spacecraft models to obtain their local models for debris processing, so as to simulate space debris targets. The latter contains a large number of stars widely distributed in space, which is used to construct the long-range space background beyond the target. Then, we used the collected data to construct a variety of small space targets through 3D modeling software, set the parameters of light reflectivity and refractive index of different materials in the three types of object models, and designed different light source positions and intensities to simulate the real space lighting environment in space. Finally, we selected multiple shooting angles, considered the different observation angles of small targets in space, and automatically shot them through the script function integrated into the software to achieve the construction of the database. Some images of the database are shown in Figure 1.

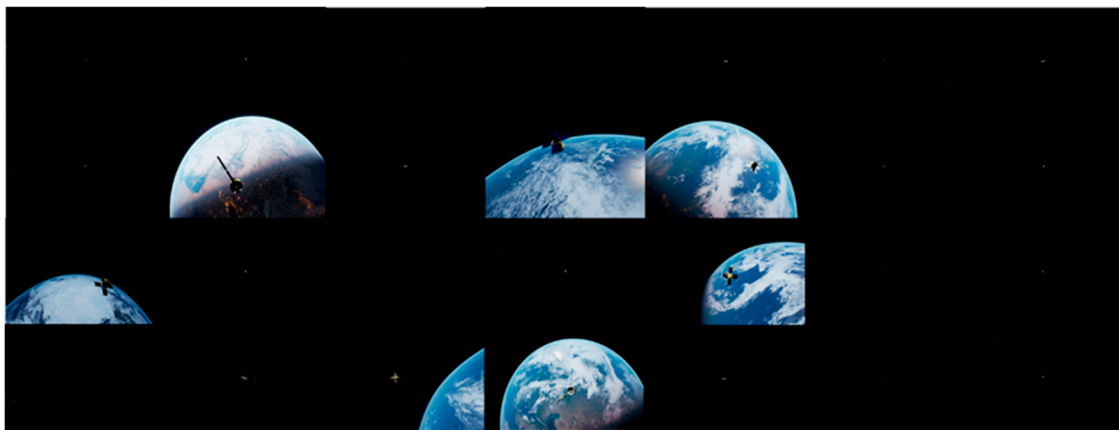


Figure 1. Images of parts of the database.

3.2. Multi-Scale Feature Fusion Pyramid Structure

The FPN [32] is a top-down feature pyramid fusing feature maps of different scales, which alleviates the contradiction between receptive field and feature extraction and enhances the performance of the detector. Subsequently, researchers proposed PANet [33] based on the FPN. PANet introduces a bottom-up sampling pathway in order to propagate the low-level feature information upwards more efficiently, so as to achieve the fusion of the low-level detail information and the high-level semantic information, and to improve the detection performance of the model. In the task of spatial small target detection, although PANet enhances the extraction of target features by introducing the strategy of top-down and bottom-up paths, there is still the problem of poor detection performance for small targets.

To address this problem, a new multi-scale feature fusion pyramid structure is proposed in this paper to enhance the detection of small targets in space, and two versions with four detectors and three detectors are provided, named MFFPN-L (Multi-scale Feature Fusion Pyramid Networks-Large) and MFFPN-S (Multi-scale Feature Fusion Pyramid Networks-Small), respectively. The feature pyramid proposed in this paper consists of three modules, i.e., the CST module, the CSE module, and the ISPP module. The pyramid structure is shown in Figure 2.

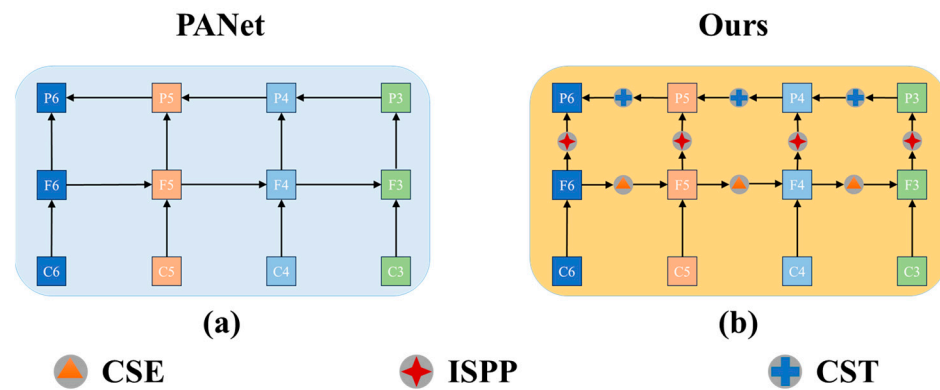


Figure 2. Multi-scale feature fusion pyramid structure. (a) is a structural diagram of PANet network. (b) is the structural diagram of our proposed multi-scale feature fusion pyramid.

In this paper, we present MFFPN-L, and the proposed multi-scale feature fusion pyramid structure is shown in Figure 2b. We were inspired by PANet to construct the feature pyramid with top-down and bottom-up features, where C3, C4, C5, and C6 are the input images that have been downsampled by the backbone network. Due to the presence of a large number of very small debris, small celestial objects, and other targets in the space small target detection task, we added a detection head for tiny targets to the traditional three detection heads to improve the detection performance for small targets. Considering the existence of interference from weak illumination and the star map background, we used the CSE module in the top-down path to improve the extraction ability of the target region through the attention mechanism and reduce the interference of illumination and star map background in target detection. Then, to improve the detection ability of small targets through multi-scale information, F3, F4, F5, and F6 were input into the ISPP module, and the improved spatial pyramid pooling was used to enhance the detection ability of small targets by fusing the features of multiple scales while increasing the sensory field acquisition. Finally, the CST module was used in the bottom-up path to extract local information in the image using the CNN, and then, positional coding was introduced by Transformer to strengthen the connection of contextual information and improve the sensory field and feature extraction ability of the model.

3.3. Converged CST Module

A convolutional neural network (CNN) is a simple and effective network for processing image tasks and has excellent performance in processing local features, but CNNs have some limitations in processing global information. The transformer model is relatively complex and has a poor ability to process local information, but it has a good ability to process global information with good long-term dependency. Inspired by ViTDet [34] and BoTNet [35], a hybrid model was constructed using a CNN and Transformer, replacing specific modules in the network and combining the advantages of both to construct a CST module. We used a Swin Transformer [36] with lower computational effort to fuse with CNN to obtain local information through the CNN, and then, the Swin Transformer captured global information and rich contextual information to improve the detection performance of small targets. As shown in the figure, each Swin Transformer contains two consecutive Swin Transformer Blocks using two modules, W-MSA (Window-based Multi-Head Self-Attention) and SW-MSA (Shifted Windows Multi-Head Self-Attention), followed by a 2-layer MLP (Multilayer Perceptron) with a GELU (Gaussian Error Linear Unit) activation function, and an LN (Layer Norm) layer is used prior to each MSA (Multi-Head Self-Attention) module and the MLP, and residual connections are used. Swin Transformer not only reduces computational complexity but also increases the ability to capture information at different scales through spatial dimensional mixing and improves detection performance for spatially small targets.

Considering that Transformer requires a large amount of computation and memory consumption, and the feature images at the end of the network have low resolution, we chose to introduce the CST module at the end of the feature pyramid. The local information in the image was extracted using the CNN, and then the linkage of contextual information was strengthened by Transformer to improve the detection ability of the model for spatially small targets, and the structure is shown in Figure 3.

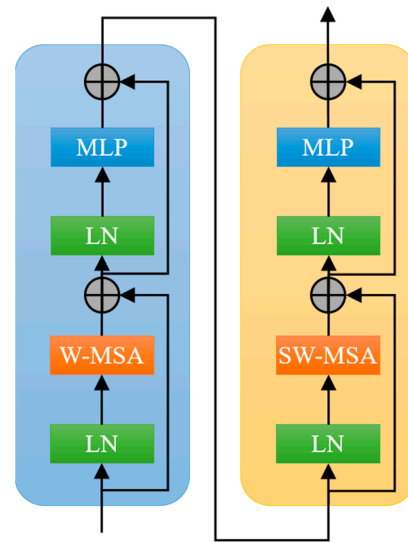


Figure 3. Transformer encoder block.

3.4. Improved SE Attention Module (CSE Module)

In the field of target detection, the SE module has gained considerable attention for its ability to adaptively extract features and has been used by a large number of networks such as SENet and MobileNet [37]. By introducing the Squeeze-and-Excitation operation to establish a relationship between the channels, the SE module is able to adaptively learn the significance of each channel's importance and adjust the weighting of the channels according to the characteristics of the detection task, thus improving the performance of the model.

Residual networks have been widely used in the field of image processing due to their powerful feature extraction ability. This module is based on a residual network, and the conventional residual network is shown in Figure 4a, which can be expressed as:

$$X_{out} = X_{in} + X_{res} \quad (1)$$

We propose the CSE module as shown in Figure 4b; first, we used lightweight grouped convolutional GConv for feature extraction of the input image and then, we used the improved SE module to obtain more important information via channel attention weighting, where the two layers of activation functions of the SE module are chosen as ReLU and H-Sigmoid, respectively. The proposed CSE module can be represented as:

$$X_{out} = w \times X'_{in} + X_{res} \quad (2)$$

where the learning weights w are calculated as:

$$r = H(R_2 * \sigma(R_1 * AAPool(GC(X_{in}))) \quad (3)$$

where $H(\cdot)$ denotes the H-Sigmoid activation function and $\sigma(\cdot)$ denotes the ReLU activation function. First, the input feature map X_{in} is extracted using grouped convolutional GConv. Then, the features are compressed via adaptive average pooling, and the weight values of

the channels R_1 and R_2 are obtained using two 1×1 Conv, which effectively improves the expressive power of the network.

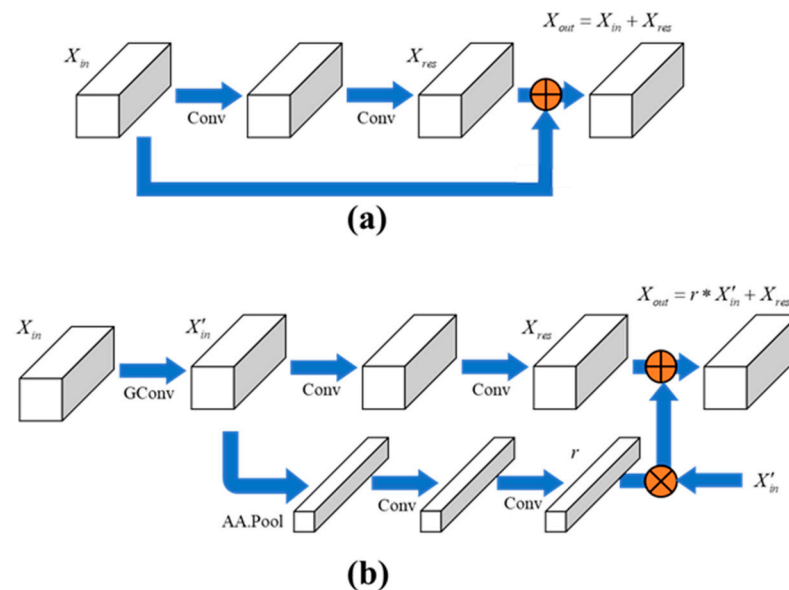


Figure 4. Improved SE attention mechanisms. (a) is a diagram of the residual network structure. (b) is the structure of the CSE module.

Attention mechanisms are difficult to use for the entire network because they increase the computation and reasoning time of the algorithm. Many existing studies show that the performance of the algorithm can be better improved when the attention mechanism is at the end of the network. Therefore, we introduced our proposed color CSE module only at the end of the network, which can better achieve accuracy–speed balance.

3.5. Improved Spatial Pyramid Pooling (ISPP Module)

Spatial pyramid pooling [38] was first proposed in 2014 and has been applied to the task of target detection, followed by a large number of research works such as those on ASPP [39], SimSPPF [40], and RFB [41]. Spatial pyramid pooling extracts and splices features together on the same image by convolving the same image with different sizes of nulls, which increases the sensory field and simultaneously improves the performance of target detection. For the task of detecting small spatial targets, we propose improved spatial feature pyramid pooling to improve the detection performance for small targets.

Our improved spatial feature pyramid pooling is shown in Figure 5. It makes the existing structure more suitable for our spatial small target detection task. Firstly, in order to obtain more information about small targets that exist in the image, we cropped the 3×3 convolution kernel in the original structure, then the activation function of ReLU was selected to improve the running speed, and finally, the size of the pooling kernel was changed from (5, 5, 5) to (3, 3, 3); the sensory field of the pooling kernel of smaller size was more suitable for the small target scale, which is conducive to the extraction of the features of the small targets and further improves the detection of small targets. The ability of small target detection was further improved.

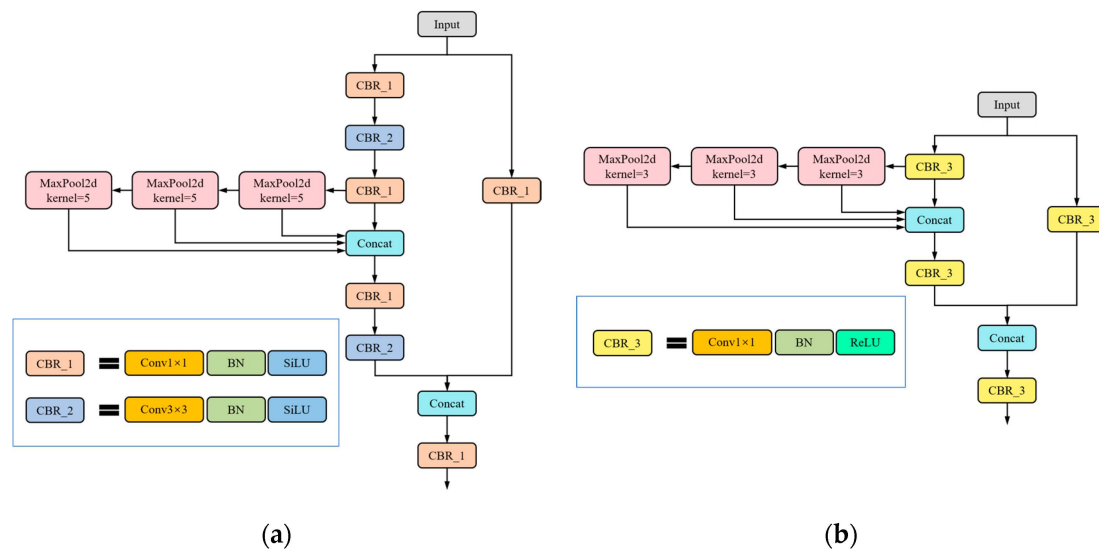


Figure 5. Improved spatial pyramid pooling module. (a) is a structural diagram of a kind of spatial pyramid pooling. (b) is the structure of the improved spatial pyramid pooling module.

4. Experimental Results and Discussion

In order to analyze the potential of the MFFPN proposed in this paper, we compared it with existing feature pyramids on a constructed spatial small target dataset. We also conducted a comprehensive ablation experiment to compare the effects of different modules on the detection performance of spatially small targets. The parameter configuration of the experimental environment is shown in Table 1.

Table 1. Parameters related to the experimental environment.

Category Name	Detail
Operating system	Windows10
CPU	Intel i7-12700 (Sanata Clara, CA, USA)
GPU	NVIDIA RTX A6000 (Santa Clara, CA, USA)
Cuda with Cudnn	v.11.7
Python	v.3.8
Pytorch	v.2.0

4.1. Experimental Setup

In this paper, in order to verify the performance of different feature pyramids, we used CSPNet [42] as the backbone algorithm and constructed a target detection algorithm using PANet, BiFPN, and MFFPN proposed in this paper for the experiments. The migration learning strategy was used in the experimental process. First, the constructed algorithm was pre-trained on the COCO dataset and the pre-trained weight parameters were obtained. Then, the weights of the algorithm were initialized using the pre-training file and trained on the constructed spatial small target dataset, i.e., the idea of migration learning was used to accelerate the training speed of the model.

(1) Experimental details

During the experiment, the batch size (the number of data samples during batch training) was determined to be 128 using automatic optimization, the image size was 640×640 , the number of epochs was 500, the initial learning rate was 0.001, the momentum factor was 0.937, and the weight decay rate was 0.0005. In order to speed up the training of the model, we used the migration learning strategy. At the same time, we used the warm-up strategy to use a smaller learning rate at the beginning of training to ensure the

stability of the model and increase the learning rate to improve the convergence speed of model training after the model was stabilized.

In this study, we used a variety of data enhancement methods to enhance the detection accuracy of the model in the presence of external disturbances. In addition to the commonly used image brightness, saturation, and noise enhancement methods, we also used geometric variation methods such as random scaling, cropping, panning, shearing, and rotation. In addition to individual image data enhancement methods, we also used mosaic data enhancement methods, where multiple images are spliced after data enhancement and the corresponding bounding boxes are changed correspondingly to obtain new images and bounding boxes. Considering the strong similarity of the star background, we used copy-pasting to enhance the data to increase the number of targets in the image and strengthen the detection performance of the model. Data enhancement can effectively extend the dataset to make the model have stronger detection performance for targets in different scenes and increase the generalization of the model for image detection in different scenes.

(2) Evaluation indicators

The performance of a model is usually evaluated using precision (P), recall (R), and mean average precision (mAP). Precision refers to the proportion of true positive samples among all samples predicted by the classifier, which can be expressed as:

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{num_{pred}} \quad (4)$$

where TP denotes the number of samples that were true positive examples, FP denotes the number of samples that were incorrectly predicted to be positive examples, and num_{pred} denotes the total number of objects recognized.

Recall is the percentage of the number of samples that are correctly predicted to be positive out of all the samples that are positive, and can be expressed as:

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{num_{sample}} \quad (5)$$

where FN denotes the number of samples that were incorrectly predicted as negative examples and num_{sample} denotes the total number of objects to be detected.

Mean average precision (mAP) is a metric for evaluating the performance of target detection algorithms. Calculating the average of average precision (AP) under each category allows for comparison and evaluation of the performance of different algorithms. AP is a measure of the algorithm's performance on different target categories by calculating the maximum precision at different recall levels.

4.2. Comparison with Existing Methods

In order to prove the effectiveness of the proposed MFFPN algorithm for spatial small target detection, in this study, we conducted comparative experimental analysis with various mainstream feature pyramids on the generated spatial small target dataset, and the comparative feature pyramids included PANet and BiFPN, which have excellent performance in detection ability. We took CSPNet as the backbone algorithm and combined PANet, and BiFPN, and this paper proposes two versions of the MFFPN to construct the target detection algorithm and conduct experiments under the same training environment and parameters, and the specific model evaluation index comparison data are shown in Table 2.

The experimental results show that the two versions of the MFFPN algorithm proposed in this paper achieve the optimum in precision, recall, and mAP, which is more suitable for the detection task of small spatial targets. This shows the effectiveness of our work in increasing the sensory field and increasing the attention mechanism, which shows that the MFFPN algorithm is more effective and advantageous in performing spatial small target detection.

Table 2. Comparison of multiple feature pyramid algorithms on spatially small target datasets.

Method	Precision	Recall	mAP@0.5
PANet	0.937	0.935	0.941
BiFPN	0.933	0.929	0.954
MFFPN-S (ours)	0.971	0.953	0.973
MFFPN-L (ours)	0.972	0.973	0.981

Our proposed MFFPN algorithm was used for the detection of small targets in space, which is mainly characterized by the detection of small targets, the presence of weak illumination and interference from the background of the star map, as well as the influence of noise in the process of image acquisition. To address the above problems, we clustered all of the labels using a clustering algorithm and regenerated a new anchor frame to make the current anchor frame more suitable for the current dataset. We also applied a variety of data enhancement methods to generate data with different scales and noise. Smaller datasets can be used, and generating data through data augmentation increases the number of datasets while mitigating overfitting in training.

The MFFPN algorithm extracts the local information of the existing targets in the image through multiple attention mechanisms of joint channel attention mechanism and multi-head attention mechanism to enhance the feature extraction ability for small targets. Meanwhile, the multi-scale object information is obtained by increasing the receptive field through improved spatial pyramid pooling, so that the network obtains different receptive fields and is able to capture information at different scales to improve the feature extraction ability for small targets. We provide two versions to improve the detection performance for small targets by utilizing the added tiny target detection head.

4.3. Ablation Experiments

In order to validate the effectiveness of each module of the MFFPN algorithm, ablation experiments were conducted for experimental validation. MFFPN-L was selected for validation, and the proposed CST module, CSE module, and ISPP module were tested. First, experiments were performed for each module separately, and the proposed modules were added in sequence for the experiments. Then, it was judged by precision, recall, and mAP metrics, and the final results are shown in Table 3.

Table 3. Comparison table of ablation experiment results.

Large	ISPP Module	CSE module	CST module	Precision	Recall	mAP@0.5
				0.937	0.935	0.941
✓				0.946	0.945	0.953
✓	✓			0.948	0.948	0.963
✓	✓	✓		0.966	0.956	0.974
	✓	✓	✓	0.971	0.953	0.973
✓	✓	✓	✓	0.972	0.973	0.981

Through comparative analysis, the MFFPN-L algorithm proposed in this paper improved by 3.5%, recall improved by 3.8%, and mAP improved by 4% relative to the initial model precision, and the experimental results show that the MFFPN-L algorithm proposed in this paper can greatly improve the model's ability to perform spatial small target feature extraction and improve the recognition accuracy of spatial small targets, which proves that the MFFPN algorithm has superior performance and is adapted to the task of spatial small target detection.

To further validate the effectiveness of the MFFPN algorithm in spatial small target detection, the test set in this dataset was visualized and some of the detection results are shown in Figure 6.



Figure 6. The effect of the MFFPN on the detection of small targets in space (the left side is the label of small targets in space, and the right side is the effect of MFFPN detection).

5. Conclusions

In this paper, we propose a multi-scale feature fusion pyramid network for spatially small target detection in complex environments. The feature pyramid is used by increasing the sensing field and using the attention mechanism in order to solve the problem of the difficult extraction of spatial small target features. The experimental results show that the algorithm in this paper is significantly better than existing methods for detecting spatial small targets in complex environments, and its average recognition accuracy for spatial small targets is higher, reaching 98.1%, which is 4% higher compared to the baseline method, and the average recognition accuracy was also improved compared to other mainstream algorithms with superior performance. In summary, the method in this paper has the potential to be widely used in spatial target research, especially for the high-precision recognition of small spatial targets and fuzzy targets.

Author Contributions: Conceptualization, H.X.; methodology, C.X.; investigation, Y.L.; writing—original draft, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zou, Y.; Zhu, Y.; Bai, Y.; Wang, L.; Jia, Y.; Shen, W.; Fan, Y.; Liu, Y.; Wang, C.; Zhang, A.; et al. Scientific objectives and payloads of Tianwen-1, China's first Mars exploration mission. *Adv. Space Res.* **2021**, *67*, 812–823. [\[CrossRef\]](#)
2. Afshar, S.; Nicholson, A.P.; Van Schaik, A.; Cohen, G. Event-based object detection and tracking for space situational awareness. *IEEE Sens. J.* **2020**, *20*, 15117–15132. [\[CrossRef\]](#)
3. Li, X.; Scheeres, D.J.; Qiao, D.; Liu, Z. Geophysical and orbital environments of asteroid 469219 2016 HO3. *Astrodynamics* **2023**, *7*, 31–50. [\[CrossRef\]](#)
4. Li, X.Y.; Scheeres, D.J. The shape and surface environment of 2016 HO3. *Icarus* **2021**, *357*, 114249. [\[CrossRef\]](#)
5. Zhou, X.; Li, X.; Huo, Z.; Meng, L.; Huang, J. Near-earth asteroid surveillance constellation in the sun-venus three-body system. *Space Sci. Technol.* **2022**, *2022*, 9835234. [\[CrossRef\]](#)
6. Wang, B.; Li, S.; Mu, J.; Hao, X.; Zhu, W.; Hu, J. Research advancements in key technologies for space-based situational awareness. *Space Sci. Technol.* **2022**, *2022*, 9802793. [\[CrossRef\]](#)
7. Uriot, T.; Izzo, D.; Simões, L.F.; Abay, R.; Einecke, N.; Rebhan, S.; Martinez-Heras, J.; Letizia, F.; Siminski, J.; Merz, K. Spacecraft collision avoidance challenge: Design and results of a machine learning competition. *Astrodynamics* **2022**, *6*, 121–140. [\[CrossRef\]](#)
8. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
9. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
10. Chen, C.P.; Li, H.; Wei, Y.; Xia, T.; Tang, Y.Y. A local contrast method for small infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 574–581. [\[CrossRef\]](#)
11. Li, Z.; Zhang, T.; Sun, M. Rapid identification and precise positioning of space targets under starry sky background. *Opt. Precis. Eng.* **2015**, *23*, 589–599.

12. Fujita, K.; Hanada, T.; Kitazawa, Y.; Kawabe, A. A debris image tracking using optical flow algorithm. *Adv. Space Res.* **2012**, *49*, 1007–1018. [\[CrossRef\]](#)
13. Xi, J.; Wen, D.; Ersoy, O.K.; Yi, H.; Yao, D.; Song, Z.; Xi, S. Space debris detection in optical image sequences. *Appl. Opt.* **2016**, *55*, 7929–7940. [\[CrossRef\]](#)
14. De Vittori, A.; Cipollone, R.; Di Lizia, P.; Massari, M. Real-time space object tracklet extraction from telescope survey images with machine learning. *Astrodynamics* **2022**, *6*, 205–218. [\[CrossRef\]](#)
15. Waisberg, E.; Ong, J.; Paladugu, P.; Kamran, S.A.; Zaman, N.; Lee, A.G.; Tavakkoli, A. Challenges of artificial intelligence in space medicine. *Space Sci. Technol.* **2022**, *2022*, 9852872. [\[CrossRef\]](#)
16. Zhou, X.; Qiao, D.; Li, X. Neural Network-Based Method for Orbit Uncertainty Propagation and Estimation. *IEEE Trans. Aerosp. Electron. Syst.* **2024**, *60*, 1176–1193. [\[CrossRef\]](#)
17. Hu, J.; Hu, Y.; Lu, X. A new method of small target detection based on the neural network. In Proceedings of the MIPPR 2017: Automatic Target Recognition and Navigation, Xiangyang, China, 28–29 October 2017; SPIE: Bellingham, WA, USA, 2018; Volume 10608, pp. 111–119.
18. González, R.E.; Munoz, R.P.; Hernández, C.A. Galaxy detection and identification using deep learning and data augmentation. *Astron. Comput.* **2018**, *25*, 103–109. [\[CrossRef\]](#)
19. Jia, P.; Liu, Q.; Sun, Y. Detection and classification of astronomical targets with deep neural networks in wide-field small aperture telescopes. *Astron. J.* **2020**, *159*, 212. [\[CrossRef\]](#)
20. Guo, X.; Chen, T.; Liu, J.; Liu, Y.; An, Q. Dim Space Target Detection via Convolutional Neural Network in Single Optical Image. *IEEE Access* **2022**, *10*, 52306–52318. [\[CrossRef\]](#)
21. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [\[CrossRef\]](#)
22. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
24. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 10347–10357.
25. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 568–578.
26. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on COMPUTER vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
27. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
29. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239.
30. Dung, H.A.; Chen, B.; Chin, T.J. A spacecraft dataset for detection, segmentation and parts recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 2012–2019.
31. Hu, Y.; Speierer, S.; Jakob, W.; Fua, P.; Salzmann, M. Wide-depth-range 6d object pose estimation in space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15870–15879.
32. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
33. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9197–9206.
34. Li, Y.; Mao, H.; Girshick, R.; He, K. Exploring plain vision transformer backbones for object detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 280–296.
35. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 16519–16529.
36. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 10012–10022.

37. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [[CrossRef](#)]
39. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [[CrossRef](#)]
40. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
41. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
42. Wang, C.Y.; Liao HY, M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.