electronics

MDPI

Article

Research on Detection and Recognition Technology of a Visible and Infrared Dim and Small Target Based on Deep Learning

Yuxing Dong *, Yan Li and Zhen Li

Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
* Correspondence: dongyuxing@126.com

Abstract: With the increasing trend towards informatization and intelligence in modern warfare, high-intensity and continuous reconnaissance activities are becoming increasingly common in battlefield environments via airborne, vehicle, UAV, satellite and other platforms. Visible and infrared images are preferred due to their high resolution, strong contrast, rich texture details and color features, and strong information expression ability. However, the quality of imaging is easily affected by environmental factors, making it crucial to quickly and accurately filter useful information from massive image data. To this end, super-resolution image preprocessing can improve the detection performance of UAV, and reduce false detection and missed detection of targets. Additionally, super-resolution reconstruction results in high-quality images that can be used to expand UAV datasets and enhance the UAV characteristics, thereby enabling the enhancement of small targets. In response to the challenge of "low-slow small" UAV targets at long distances, we propose a multi-scale fusion super-resolution reconstruction (MFSRCNN) algorithm based on the fast super-resolution reconstruction (FSRCNN) algorithm and multi-scale fusion. Our experiments confirm the feasibility of the algorithm in reconstructing detailed information of the UAV target. On average, the MFSRCNN reconstruction time is 0.028 s, with the average confidence before and after reconstruction being 80.73% and 86.59%, respectively, resulting in an average increase of 6.72%.

Keywords: object detection; deep learning; visible light target; infrared target

Citation: Dong, Y.; Li, Y.; Li, Z. Research on Detection and Recognition Technology of a Visible and Infrared Dim and Small Target Based on Deep Learning. *Electronics* **2023**, *12*, 1732. https://doi.org/10.3390/electronics12071732

Academic Editor: Xin Ning

Received: 7 February 2023
Revised: 22 February 2023
Accepted: 22 February 2023
Published: 5 April 2023

1. Introduction

With the improvement of information and intelligence of war, the war mode becomes increasingly complex. The accurate perception of battlefield situation information is an essential prerequisite for guiding military operations. Among them, the use of vehicle-mounted, airborne, unmanned aerial vehicle, satellite and other platforms is an important technical means to obtain battlefield situation information [1], which can provide a large number of accurate, reliable and intuitive battlefield image information for decision makers [1]. On the one hand, the improvement of war informatization provides decision makers with more means to obtain information, but also produces massive data. Traditional manual processing methods have been difficult to meet the needs of modern warfare [2], so how to quickly find useful information from massive data has become an urgent problem to be solved in the battlefield information perception system [2]. On the other hand, in order to adapt to the intelligent development direction of future war, it is necessary to give intelligent weapons the ability to make autonomous decisions, and how to independently obtain the information of hostile targets is one of the key technologies for the development of intelligent weapons.

As an important part of a battlefield information perception system, image data is intuitive and reliable. Visible and infrared images are the main sources of image data, which are simple to obtain and widely used. In general, visible images have a high resolution, strong contrast, rich texture details and color features, strong information expression ability, and are easy to understand. However, the imaging quality of visible images is easily

affected by the environment, and the imaging effect is extremely poor in environmental conditions such as backlight, rain, smoke and so on. The infrared image is the embodiment of the infrared radiation intensity of the object, and it still has stable and reliable imaging quality in the case of the above visible image imaging failure. It is more suitable for the complex battlefield environment, and has the ability to work stably in all weather conditions. A visible image and an infrared image exhibit a good complementary effect, and it is of great military significance to obtain hostile target information through these two kinds of image data.

In April 2017, the United States developed the MLIDS integrated defense system [3]. As shown in Figure 1, by installing photoelectric infrared sensors and electronic jamming systems, radar detection and kinetic energy weapons on two mine-proof and anti-ambush vehicles (M-ATV), the UAV can be identified and tracked, respectively, and kinetic energy weapons can be used to strike targets [3]. The UAVX anti-UAV system [4] developed by the Black Monitor Company in the United States, as shown in Figure 2, is composed of an infrared camera with n 8× zoom, a visible light camera with a 6× optical zoom, and a radar with a 500 m detection range [4]. The neural network is deployed on a mobile computer with A15CPU and 192 cores of CUDA GPU to process and classify radar data. Effective UAV target detection and recognition can be achieved within 500 m, and zoom infrared and visible light cameras are used for target tracking. The same American surveillance and control anti-UAV system also includes the use of the optical and infrared detection NOAS system; the TCUT system integrating radar, optical and infrared detection [5] and so on [5].

Figure 1. MLIDS integrated defense system.

Figure 2. UAVX anti-UAV system.

In recent years, with the help of the progress of hardware technology and the proposal of related theories, such as the convolutional neural network, object detection technology based on deep learning has developed rapidly and has been successfully applied to fields such as face detection, autonomous driving, pedestrian recognition, medical image analysis,

security video detection and so on. However, in the military field, due to the requirements of security and confidentiality, the object detection technology based on deep learning has been developed rapidly. The deep learning detection technology for military targets is still under development. Military scientific research institutions in various countries, represented by DARPA, foresee the huge application value of deep learning technology in the military field, and attach great importance to the transformation application and innovative development of deep learning technology in the military field [6]. Among them, battlefield target detection technology based on deep learning is one of the research focuses [6].

Therefore, the deep learning detection technology of visible and infrared targets in military application scenarios can not only assist decision makers to quickly screen information, but it can also be one of the key technologies in the process of intelligent weapons and equipment, which has extremely important military application value. In modern war, soldiers and armored vehicles are the two main types of combat units. Based on these two types of targets, this paper carries out research on target detection technology based on deep learning, which realizes the accurate identification and detection of visible light targets under appropriate lighting conditions, as well as the accurate identification and detection of infrared targets under weak light conditions, such as dark and dim conditions.

2. Related Work

Object detection is one of the four research topics in the field of computer vision. Its significance is to locate and classify the object of interest in the image, so as to provide accurate and reliable object information for subsequent applications. In recent years, thanks to the rapid development of deep learning technology, object detection methods based on deep learning technology have surpassed traditional methods and become the main research direction of the development of object detection technology.

Before 2012, object detection technology was in the traditional development stage, and object detection algorithms mainly relied on statistical analysis and manually designed features. However, the manually designed features had limitations, feature design became more and more complex, and the development of object detection technology entered a bottleneck period. Krizhevsky et al. [7] proposed the AlexNet neural network, which was the first to use convolutional neural network to complete the target classification task, and won the ImageNet [8] challenge in the same year with an overwhelming advantage. The successful application of convolutional neural networks has widely attracted the attention of the academic community [7,8]. In 2014, Girshick et al. [9] proposed the R-CNN algorithm, which successfully applied the convolutional neural network to the object detection task, marking that the object detection algorithm officially embarked on the development route based on deep learning technology, and the object detection technology developed rapidly [9].

In 2016, Redmon et al. [10] proposed the first one-stage object detection algorithm YOLO. YOLO takes the whole image as the network input, extracts the global feature map of the image, and directly regress the position and category of the object based on the global feature map 10. Because YOLO has only one neural network, the training and detection process is easier, which greatly improves the detection speed. The object detection algorithm has the real real-time detection ability, but the detection accuracy is lower than the two-stage object detection algorithm, and the detection effect of small objects is not good. The YOLOv2 [11] algorithm focuses on solving the problems of low recall rate and large positioning error of the YOLO algorithm 11. Five anchor boxes of different sizes are designed to greatly improve the recall rate of the algorithm. The YOLOv3 [12] algorithm proposes a new backbone network Darknet-53, and uses a convolution kernel with twice the step size to replace the pooling layer to realize the structure of a fully convolutional neural network 12. Due to the excellent performance of the YOLOv3 algorithm, the subsequent improvements are based on YOLOv3, and the improvement direction is mainly in the network structure, loss function, feature fusion mode, etc. For example, ASFF [13]

improves the feature fusion method of YOLOv3, and proposes an adaptive spatial feature fusion method, which improves the scale invariance of features and effectively improves the detection accuracy without affecting the detection speed [13]. In 2020, the YOLO series of algorithms ushered in a new version of YOLOv4 [14], which widely draws on a variety of object detection improvement techniques proposed in recent years, further enhancing the detection accuracy [14].

Deep learning object detection technology has developed rapidly, but the mainstream research is still limited to the visible light field. The imaging quality of visible images is easily affected by environmental changes, and the imaging effect is extremely poor in weak light environments, such as night, haze, rain and so on. The infrared image is the reflection of the temperature difference of the object, and the imaging effect is still reliable in the above environment and has the ability to work all day. Thanks to the progress of infrared imaging technology, the application of infrared imaging is increasingly common. Target detection technology in infrared scenarios has become one of the key research directions of scholars at home and abroad in recent years. However, due to the shortcomings of infrared images, such as low resolution, low contrast, more noise and fewer features, new challenges are posed to target detection technology in infrared scenes.

Ghose et al. [15] proposed an infrared pedestrian target detection technology using the saliency feature map. The saliency feature map of the infrared image is extracted by the saliency algorithm, and then fused with the original infrared image for the training and testing of the Faster R-CNN network. The fusion method effectively improves the detection accuracy of infrared pedestrian targets [15]. Laixiang et al. [16] proposed an improved infrared target detection method based on the ZFNet network. For the feature analysis of infrared targets, a spatial transformation network and dropout [17] layer were introduced on the basis of ZFNet, which effectively improved the recognition accuracy of the model. Srivastava et al. [17] and Hou et al. [18] proposed a detection network RISTDnet for infrared small targets, which uses a convolutional neural network and hand-designed feature methods to extract feature maps of infrared images, and detects small infrared targets based on target and background possibility feature maps, which has good detection accuracy for small targets with a low signal-to-noise ratio in complex backgrounds [18]. Wang et al. [19] designed a feature extraction backbone network MNET dedicated to infrared small targets. In order to retain the structural features of infrared small targets as much as possible, the network only has three down sampling operations, and uses a dense connection structure to integrate the position information of shallow features and the semantic information of deep features to achieve a better target localization and classification effect. The channel attention mechanism is introduced to improve the importance of useful features, which effectively improves the detection effect of infrared small targets [19]. For the detection of aerial infrared targets, Jiangrong et al. [20] proposed an infrared target detection method based on single shot multibox detector (SSD), which uses pooling and transposed convolution operation to complete the bidirectional fusion of features, achieving better feature extraction ability. By adding feature enhancement branches and more prediction frames in the shallow layer of the network, the detection accuracy of infrared small targets is effectively improved [20].

In general, target detection technology in infrared scenes has achieved certain results in recent years, but most of the methods directly transfer to the target detection algorithm designed for visible light scenes, and the lack of algorithm design for infrared image characteristics. The existing open-source datasets are mainly concentrated in the visible light field, and there are few open-source datasets for infrared targets. Due to the uneven quality of infrared devices, the infrared images between different datasets are quite different, which affects the development of target detection technology in infrared scenes.

3. Multi-Scale Fusion Super-Resolution Reconstruction Algorithm for UAV Detection

In UAV detection, it is necessary to ensure that the early warning time is advanced to provide sufficient time conditions for subsequent operations. The detection system needs to detect and identify the target when the target is at a long distance. However, the proportion of long-distance non-cooperative UAV in the detection field of view is very small, the imaging pixels are limited, the target characteristics are seriously degraded, and it is easy to be interfered by complex background. It is difficult to distinguish low and slow small targets from airborne objects such as kites and hot air balloons, which affects the effect of target recognition and produces certain misjudgment and missed detection. In order to enhance the characteristics of UAV targets, this chapter uses the image super-resolution reconstruction method to reconstruct the low-resolution image with UAV targets to obtain a high-resolution image of the same scene. The prior knowledge of the UAV contour and the single frame of the UAV image in the video can be used to recover the high-resolution detailed features of the target. Based on the small target detection of UAV, a new network architecture is proposed on the basis of FSRCNN algorithm, and the network structure of multi-scale fusion super-resolution algorithm is constructed.

3.1. Principle Framework

Multi-scale fusion super-resolution reconstruction convolutional neural networks (MFSRCNN) is based on the consideration that the long-range UAV target is not obvious. In the case of a small number of pixels and unclear contours, based on the network structure of the FSRCNN algorithm, combined with the idea of HrNet, the algorithm is obtained by a parallel connection of four subnets with different resolutions and multiple multi-scale fusion.

As shown in Figure 3, the network structure is mainly composed of four branches with different resolutions. Feature maps of corresponding resolution images are obtained through feature extraction and nonlinear transformation, respectively. Feature fusion between adjacent resolution feature maps integrates subnetwork image information, and then the target UAV image is obtained through image reconstruction. In this network structure, p is 168, q is 36, and m is 12.

3.1.1. Feature Extraction

The reconstruction from LR image to HR image needs to be achieved by upsampling. MFSRCNN upsampling is similar to FSRCNN with the deconvolution operation. Deconvolution is a type of zero-padding operation on the input image followed by convolution.

The operation of the parameters needs to be learned. The calculation formula of deconvolution is shown in Equation (1).

$$O = s\bullet(i-1) + k - 2p \tag{1}$$

where s stands for step size, k for kernel size and p for padding.

All the upsampling uses deconvolution, which can increase the receptive field, improve the image quality, and transform the reconstruction process into an end-to-end autonomous learning process.

MFSRCNN uses Conv and Deconv to upsample and downsample the image, respectively, to obtain feature maps of four resolution sizes, which are used as the input of four parallel network branches. Parallel networks are named subnetwork 1, subnetwork 2, subnetwork 3 and subnetwork 4 from top to bottom. Convolution performs feature extraction at four scales while changing the image size. The deconvolution in the upsampling process is composed of multiple serial deconvolution (step size is 2) when the upsampling is greater than 2. See Figure 4 for the detailed process.

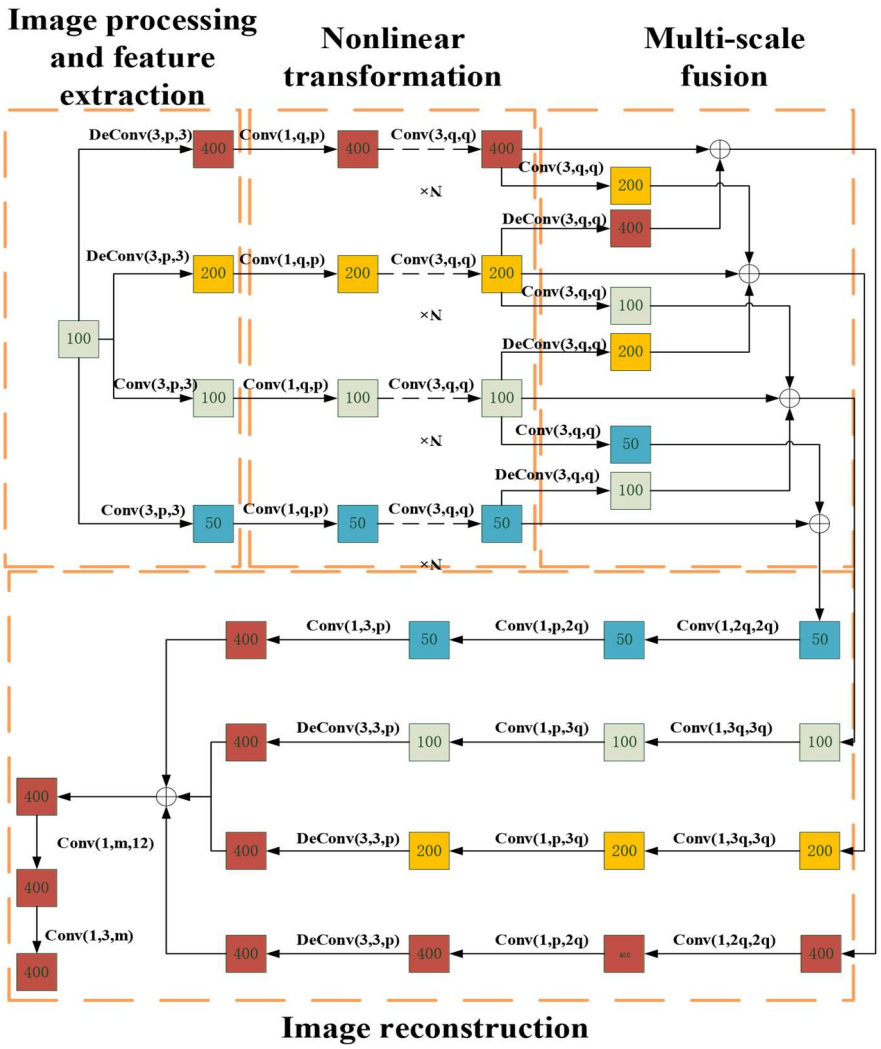


Figure 3. Network structure diagram of the MFSRCNN algorithm.

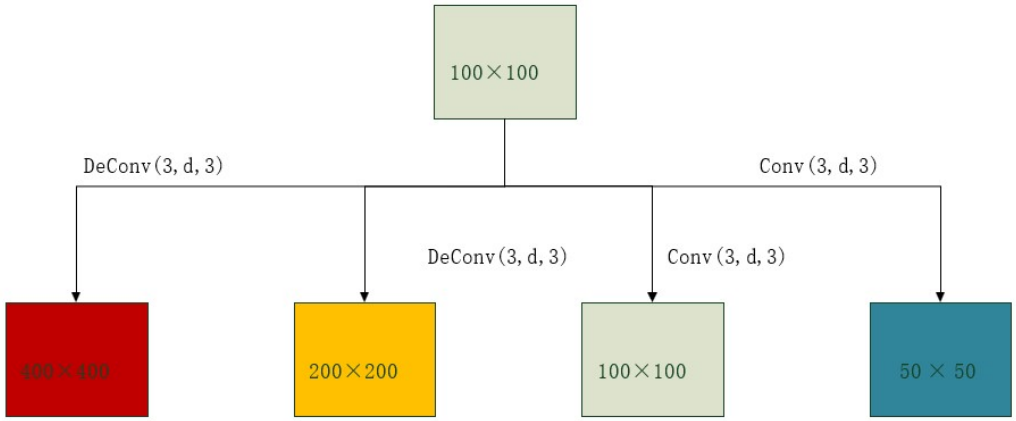


Figure 4. Deconvolution and feature extraction structure diagram.

Taking subnetwork 1 as an example, 168 convolution kernels are used for feature extraction, and Figure 5 shows the results of upsampling and feature extraction with the first 10 convolution kernels. The deconvolved image is enlarged to 400 × 400, different convolution kernels extract different features, and some convolution kernels are sensitive to the UAV contour, such as convolution kernels 1 and 7. Some convolution kernels are

more sensitive to UAV internal features such as 4 and 5. The 168 convolution kernels jointly extract UAV features from contour, color, detail and other aspects.

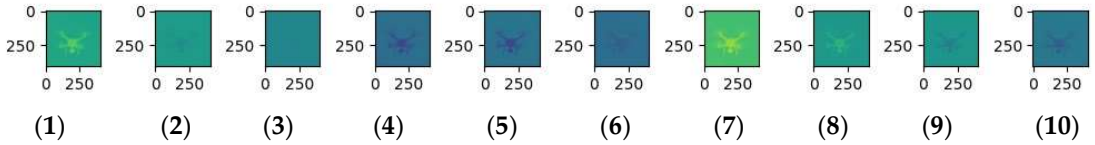


Figure 5. Partial convolution kernel output in subnetwork 1.

3.1.2. Nonlinear Transformations

On the basis of the FSRCNN nonlinear transformation operation, four subnetworks are parallelized by 1 × 1 convolution kernel to reduce the dimension from 168 to 36, which can greatly reduce the amount of computation. Then, two 3 × 3 convolutional layers are used for feature mapping in parallel. Figure 6 shows the structure diagram of nonlinear transformation.

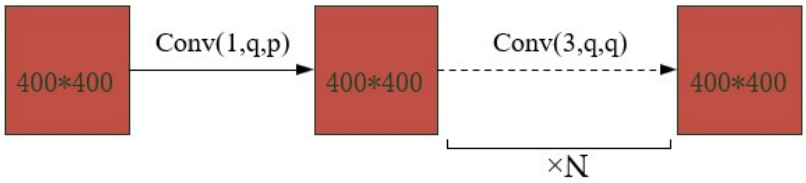


Figure 6. Structure diagram of nonlinear transformation.

3.1.3. Multi-Scale Fusion

Multi-scale fusion fuses feature maps of different scales to enhance image details, which is conducive to improving the accuracy of super-resolution reconstruction [21]. Images with different resolutions are directly input into different subnetworks, and then feature fusion is performed. The detailed fusion steps are shown in Figure 7. The feature maps are upsampled and downsampled, respectively, and fused with the corresponding size feature maps.

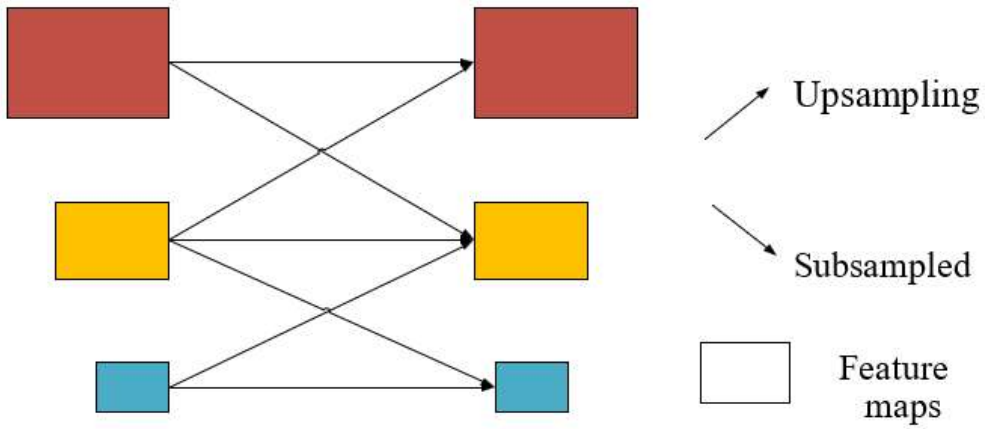


Figure 7. Schematic diagram of multi-scale fusion.

As shown in Figure 8, the feature maps of different subnetworks are, respectively, upsampled and downsampled to adjust the size of the feature maps of adjacent subnetworks. Each subnetwork is fused with feature information from adjacent parallel subnetworks. Through multiple multi-scale feature fusion of the feature map, the subnetwork information is constantly exchanged and fused, which is conducive to the extraction of target detail features in super-resolution reconstruction.

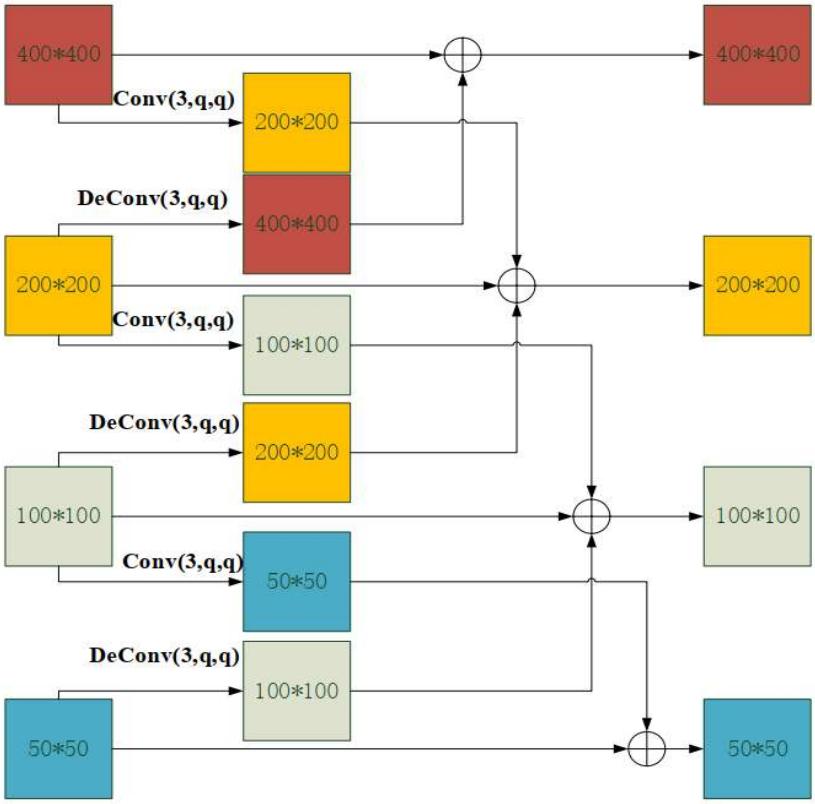


Figure 8. Structure diagram of multi-scale fusion.

3.1.4. Image Reconstruction

The feature fusion results are extracted separately, and the dimension is expanded by a 1 × 1 convolution kernel to solve the problem of the poor reconstruction effect of low-dimensional feature maps. All four parallel subnetworks are upsampled to the target size for feature fusion and feature extraction to obtain the final super-resolution image, as shown in Figure 9.

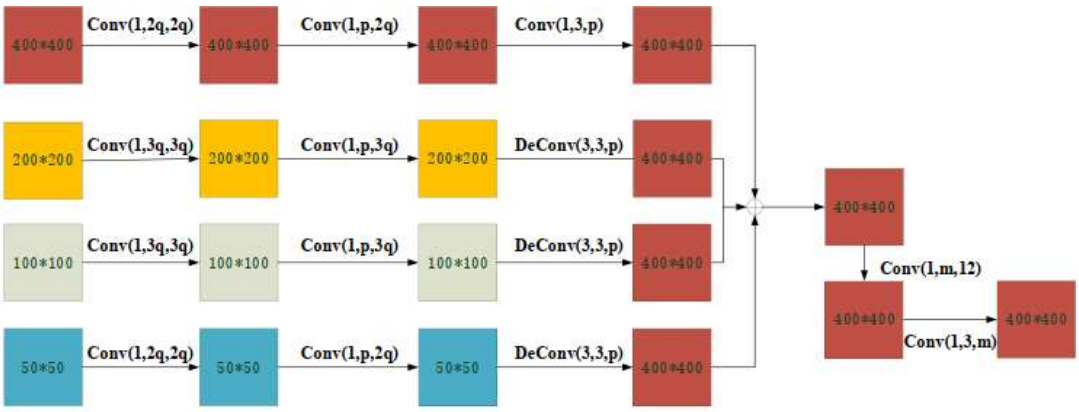


Figure 9. Image reconstruction structure diagram.

The fused feature map contains three-channel feature maps from four network branches, and the feature map 10 is obtained after the fused result is sliced.

The fusion result shows that the feature map image in subnetwork 1 has a clearer contour, richer details, and contains more detail and contour information. Figure 10 shows the feature map of feature fusion result. The smaller the original input of the network is, the less the contour and detail features are in the output, and the more abundant the low-frequency information is. The fused image can be effectively compatible with high

frequency information and low frequency information. Figure 11 is a grayscale image of the feature extraction of 12 convolutional kernels. Some kernels extract information such as contour information, e.g., convolution kernel 9, and some kernels extract low-frequency information inside the UAV (excluding texture features), such as convolution kernel 1, 2, etc. The 12 convolution kernels are extracted together to form the feature extraction part.

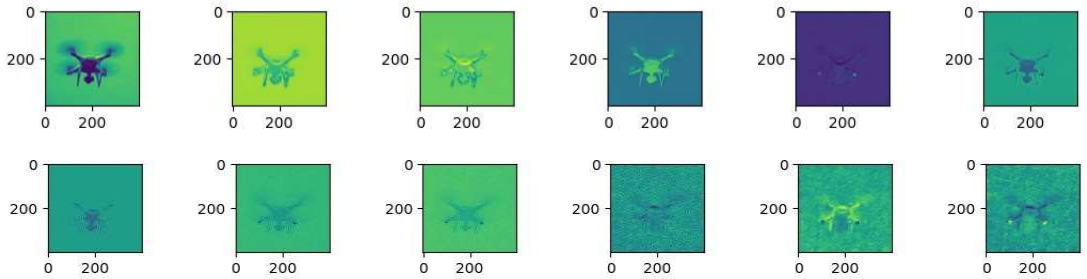


Figure 10. Feature map of feature fusion result.

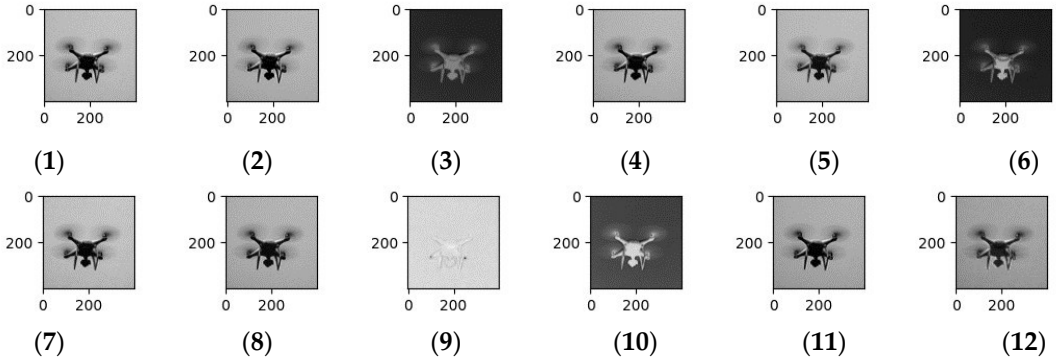


Figure 11. Feature extraction result feature map.

The Relu function is used to activate the convolution layer (except deconvolution) to accelerate the training process. At the same time, the large convolution kernels in FSRCNN are replaced with multiple 3 × 3 small convolution kernels to ensure the reconstruction quality. In the network, zero-padding is used to recover the image information to avoid truncation error. Xavier was used to initialize the weight w in the convolutional layer, and normal distribution was used to initialize the deconvolution layer; all parameters b were initialized to 0.

3.2. Simulation Training and Results

3.2.1. Model Training

The data samples used in this paper are all acquired experimentally. The sample original image size is 1920 × 1080.

Camera data: Sensor: 1/2.8 inch CMOS;

Focal length: 8.1–310 mm (F1.8–F5.6).

The number of samples collected in the experiment is limited, and data augmentation is used to expand the training sample set. Through affine transformation, gamma transformation and some trimming work, the dataset of the original samples of UAV is expanded, and finally about 1650 sample data containing UAV are obtained.

After data augmentation, UAV data samples are still limited. Firstly, the MFSRCNN network is used to train the COCO dataset to obtain the pre-trained model, which improves the feature extraction ability of the model and can effectively avoid overfitting [22]. Then, fine-tune is used to train the UAV data to achieve super-resolution reconstruction of UAV images.

The MSE loss function is used for training and the Adam optimization algorithm is used.

The input data of the UAV is 100 × 100, output super-resolution size is 400 × 400, and the training set and test set are 4:1. Because of the limited GPU memory, the batch size is set to 8, and the learning rate is set to 0.0001; the learning rate undergoes a certain attenuation during the training process. The loss curves obtained after pre-training and training with UAV data are shown in Figure 12.

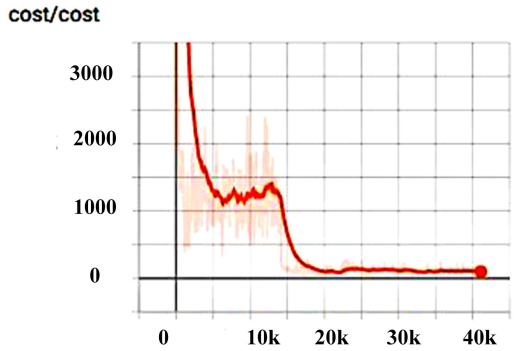
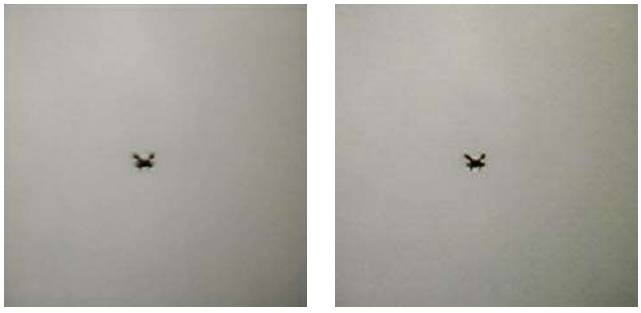


Figure 12. Loss change curve during training.

Among them, the first 13,000 steps are to pre-train the COCO dataset and enhance its feature extraction ability. After the curve is almost smooth, the transfer learning of UAV data is started. The learning rate is set to 0.0001 to continue training, and the loss starts to level off at about 30,000. As shown in Figure 13, training can effectively enhance the reconstruction details and improve the reconstruction effect to a certain extent.



(**a**) (**b**)

Figure 13. Comparison of super-resolution reconstruction results with and without pre-training. (**a**) Results without pretraining and (**b**) results with pretraining.

3.2.2. Experimental Results

After training, the UAV super-resolution reconstruction results are obtained, and compared with the bicubic interpolation and FSRCNN reconstruction results. Bicubic interpolation, FSRCNN and MFSRCNN all have certain reconstruction effects on UAV blurred images. Among them, the UAV target reconstructed by bicubic interpolation is relatively fuzzy, losing the contour features and detail features of the UAV, and it is just a simple enlargement of the image, and there will be a certain degree of black edge (which can be removed by algorithm improvement). It is not conducive to the later UAV identification. FSRCNN has a good reconstruction effect and clearly shows the contour features of the UAV, but the background noise in the reconstruction results is large. Through comparison, it is found that, on the one hand, MFSRCNN can significantly improve the reconstruction effect of UAV images and it can effectively suppress background noise and improve the signal-to-noise ratio of the image. On the other hand, by using multiple multi-scale fusion and COCO dataset pre-training, the features of the UAV can be better extracted in super-resolution reconstruction, the detailed features can be always retained in repeated multi-scale fusion, and the detailed information of the UAV can be effectively reconstructed.

4. Analysis of Experimental Results

4.1. Comparative Analysis of Models

The MFSRCNN algorithm and other classical super-resolution reconstruction algorithms were tested on the UAV test set images, and the PSNR and SSIM under different algorithms were statistically calculated, and Table 1 was obtained. Because the Loss function of ESRGAN adopts Perceptual Loss, although the reconstructed visual effect is good, it is not suitable for PSNR measurement, so Table 1 is not included in the ESRGAN comparison.

Table 1. Comparison of PSNR and SSIM and parameters reconstructed by different algorithms for the UAV test set.

Method	Params	PSNR	SSIM
Bicubic	15 K	17.2557	0.6607
FSRCNN	63 K	22.4835	0.7503
VDSR	665 K	26.0151	0.7698
LapSRN	813 K	26.2043	0.7726
DRRN	297 K	27.6547	0.7789
SRResNet	1 518 K	31.9431	0.7974
MFSRCNN	406 K	31.8800	0.7961

The overall framework and architecture of MFSRCNN are very simple, but it is more effective than many previous models, especially for the problem of UAV super-resolution reconstruction of small targets. MFSRCNN achieves a good balance between model size and reconstruction performance. Although MFSRCNN has more parameters than FSRCNN, DRRN and other lightweight network models, its effect on the UAV test set is substantially different to other network models. Although SRResNet has a good reconstruction effect, its network is too complex, and its network parameters are 3–4 times that of the MFSRCNN model, which requires more network training and UAV super-resolution reconstruction time. According to Table 1, the performance and parameter comparison results of MFSRCNN and other algorithms are obtained, as shown in Figure 14.

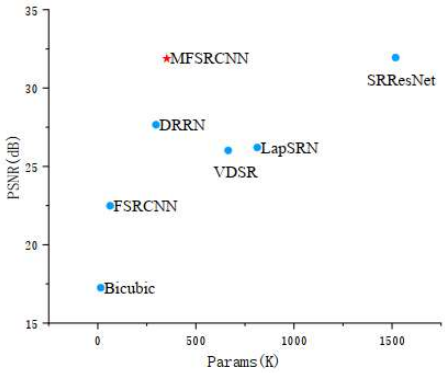


Figure 14. MFSRCNN performance and parameters compared to other advanced lightweight networks.

The experimental results show that the average reconstruction time of MFSRCNN is 0.028 s under GPU (NVIDIA GeForce GTX 1080 Ti). In the case of a small input image, it meets the requirements of real-time detection of UAV.

4.2. Analysis of Super-Resolution Reconstruction Effect

In order to measure the super-resolution reconstruction effect of UAV images, part of the images in the UAV dataset before and after reconstruction are input into the YOLOv3 network for UAV identification in turn, and the partial results are shown in Table 2.

Table 2. Detection results by YOLO network before and after UAV super-resolution reconstruction.

Serial Number	Before	After	Ascension	Serial Number	Before	After	Ascension	Serial Number	Before	After	Ascension
1	0.82	0.86	4.88%	11	0.7	0.73	4.29%	21	0.85	0.91	7.06%
2	0.95	0.97	2.11%	12	0.62	0.67	8.06%	22	0.72	0.68	−5.56%
3	0.76	0.79	3.95%	13	0.73	0.78	6.85%	23	0.5	0.52	4.00%
4	0.86	0.9	4.65%	14	0.63	0.69	9.52%	24	0.56	0.65	16.07%
5	0.63	0.65	3.17%	15	0.68	0.77	13.24%	25	0.96	0.97	1.04%
6	0	0	0.00%	16	0.93	0.95	2.15%	26	0.96	0.98	2.08%
7	0.71	0.77	8.45%	17	0.67	0.98	46.27%	27	0.75	0.91	21.33%
8	0.91	0.92	1.09%	18	0.8	0.84	5.00%	28	0.97	0.99	2.06%
9	0.93	0.94	1.08%	19	0.83	0.83	0.00%	29	0	0.41	∞
10	0.45	0.48	6.67%	20	0.61	0.73	19.67%	30	0.88	0.97	10.23%

It can be seen that super-resolution reconstruction has improved the detection accuracy of UAV to a certain extent. The average confidence before reconstruction is 80.73%, and the average confidence after reconstruction is 86.59%, with an average increase of 6.72%. For the case where the low-resolution UAV image is obvious and the score is large before reconstruction, the detection improvement effect is not obvious, and the confidence is increased by about 2%. For the case where the LR UAV image is not obvious and the score is relatively small before reconstruction, the reconstruction can effectively improve the probability of UAV classification prediction and the accuracy of boundary box localization. In the case of some UAV image contour features where are not obvious and feature degradation is serious, the LR image can not detect the UAV, and the SR image can detect the UAV, which improves the detection accuracy of the UAV. In a few cases, there will be no improvement or even degradation of UAV detection, and the super-resolution model needs to be further improved. It can confirm the effectiveness of the MFSRCNN algorithm and has practical application value.

5. Conclusions

Given the challenges of detecting long-distance UAV targets with a small proportion of pixels and unobvious contour features, we propose a novel MFSRCNN algorithm that combines the FSRCNN algorithm with the HrNet network. The MFSRCNN algorithm leverages multi-scale fusion to preserve the high resolution and reduce the loss of UAV information during reconstruction, resulting in improved target detection and reduced noise artifacts. Our experimental results demonstrate that the algorithm effectively reconstructs detailed information of the UAV target, with an average time of 0.028 s and an average confidence increase of 6.72%. With fewer network parameters, the MFSRCNN algorithm achieves an optimal balance between performance and model size, meeting the real-time demands of UAV detection. Our verification of the detection algorithm shows that the reconstruction significantly improves the confidence of the UAV target and provides a solid foundation for small target detection.

Author Contributions: Conceptualization, Z.L.; Data curation, Z.L.; Formal analysis, Y.L.; Investigation, Y.D.; Methodology, Y.D.; Writing—original draft, Y.D.; Writing—review and editing, Y.D. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

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

References

1. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9627–9636.
2. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 2918–2928.
3. Wei, S.; Kang, J.U. Optical flow optical coherence tomography for determining accurate velocity fields. *Opt. Express.* **2020**, *28*, 25502. [CrossRef] [PubMed]
4. Xue, Q.; Zhu, Y.; Wang, J. Joint Distribution Estimation and Navie Bayes Classification Under Local Differential Privacy. *IEEE Trans. Emerg. Top. Comput.* **2021**, *9*, 2053–2063. [CrossRef]
5. Li, L.; Wei, Y.; Zhang, Y. The development of anti-UAV technical equipment of the U.S. armed forces. *Aerosp. Electron. Warf.* **2017**, *33*, 60–64.
6. Hao, J.; Luo, S.; Pan, L. Computer-aided intelligent design using deep multi-objective cooperative optimization algorithm. *Future Gener. Comput. Syst.* **2021**, *124*, 49–53. [CrossRef]
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
8. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
9. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Tern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
10. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–27 June 2016; pp. 779–788.
11. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
12. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
13. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
14. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934,.
15. Ghose, D.; Desai, S.M.; Bhattacharya, S.; Chakraborty, D.; Fiterau, M.; Rahman, T. Pedestrian Detection in Thermal Images using Saliency Maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 1–10.
16. Xu, L.; Lu, G.; Liu, S.; Xiu, C.L. Research on Infrared Target Recognition Method Based on Improved CNN. *Fire Control. Command. Control.* **2020**, *45*, 136–141.
17. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
18. Wang, K.; Li, S.; Niu, S.; Zhang, K. RISTDnet: Robust Infrared Small Target Detection Network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5.
19. Wang, K.; Li, S.; Niu, S.; Zhao, Y.; Zheng, H.; Zhang, W. Detection of Infrared Small Targets Using Feature Fusion Convolutional Network. *IEEE Access* **2019**, *7*, 146081–146092. [CrossRef]
20. Xie, J.; Li, F.; Wei, H. Enhanced single—shot multi—frame detector method for aerial infrared target detection. *J. Opt.* **2019**, *39*, 223–231.
21. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. pp. 5693–5703.
22. Ribeiro, E.; Uhl, A.; Wimmer, G.; Häfner, M. Exploring Deep Learning and Transfer Learning for Colonic Polyp Classification. *Comput. Math. Methods Med.* **2016**, *2016*, 6584725. [CrossRef] [PubMed]