Article

PFFNET: A Fast Progressive Feature Fusion Network for Detecting Drones in Infrared Images

Ziqiang Han [1], Cong Zhang [1,*], Hengzhen Feng [2], Mingkai Yue [1], and Kangnan Quan [1]

[1] School of Equipment Engineering, Shenyang Ligong University, Shenyang, Liaoning 110159, China
[2] Science and Technology on Electromechanical Dynamic Control Laboratory, Beijing Institute of Technology, Beijing 100081, China
* Correspondence: zcbxl@sylu.edu.cn

Abstract: The rampant misuse of drones poses a serious threat to national security and human life. Currently, the CNN method has been widely used in drone detection. However, there is a challenge that traditional CNN cannot cope with, which is small drone targets often have reduced amplitude or even lost features in infrared images. This paper proposes a Progressive Feature Fusion Network (PFFNet) that gradually increases the response amplitude of the target in the deep network. The Feature Selection Model (FSM) is designed to improve the utilisation of the output coding graph and enhance the feature representation of the target in the network. A lightweight segmentation head is also designed to achieve progressive feature fusion with multi-layer outputs. Experimental results show that the proposed algorithm can achieve low duration and high accuracy in drone target detection. On the public dataset, the IoU is improved by 2.53% and the detection time is reduced by 81.03%.

Keywords: Infrared small targets; Counter-drones; Progressive fusion; Lightweight network

1. Introduction

Various types of small unmanned aerial vehicles pose a serious threat to infrastructure, hardware and people [1]. At the same time, the accurate detection of drone targets in low resolution, visually blurred infrared images is a challenging task. There are two main problems:

1) **The influence of the target itself**: due to the long imaging distance, infrared targets are generally small, with only a few to several tens of pixels in the image. In addition, infrared targets usually have a low signal-to-noise ratio (SCR) and are easily submerged in strong noise and cluttered backgrounds. Therefore, the radiation intensity of the target is lower and it lacks significant morphological features, making target detection in infrared images difficult.

2) **The contradiction between the target and the detection algorithm**: compared to targets in visible light images, targets in infrared images present more challenging problems. Such as the lack of shape and texture features, which leads to the weakening or even loss of high-frequency amplitude of small targets after filtering and convolution calculations. Besides, although building shallow networks can improve performance in deep learning algorithms, the contradiction between advanced semantic features and high resolution still cannot be resolved.

Overall, there are too many negative samples in the image due to the large variation in target size and the extremely low percentage of pixels in infrared images, resulting in the loss of most of the available information during algorithm operation [2]. In addition, most negative samples are easily classified, which makes it difficult for the algorithm to optimize in the expected direction. Therefore, the nets designed for normal objects is hardly use to detect small infrared targets.

To detect small infrared targets, researchers have proposed many traditional methods over the past few decades. The traditional detection method involves implementing SIRST (Single-frame InfraRed Small Target) detection by calculating the non-coherence between target and background. Typical methods include filter-based methods [3-5], which can only suppress uniform and smooth

 doi:10.20944/preprints202305.0100.v1

background noise, resulting in high false alarm rates and unstable performance for complex backgrounds. The HVS method [6-9] uses the ratio of gray values between each pixel position and its neighbouring region as an enhancement factor, which can effectively enhance the real target. However, it cannot effectively suppress the background noise. Methods based on low-rank representation [10-12] can adapt to infrared images with low SCR ratios. However, in complex backgrounds, there is still a high false alarm rate for small and shape-varying targets. Most traditional methods heavily rely on manual features. These methods are simple calculation and do not require training or learning. However, designing hand-crafted features and tuning hyperparameters require expert knowledge and a significant amount of engineering efforts.

With the development of CNN methods, more data-driven methods are being applied to infrared small target detection [13-16]. Data-driven methods are suitable for more complex real scenarios, and are less affected by target size, shape, and background changes. These methods require a large amount of data to demonstrate strong model fitting ability and have achieved better detection performance than traditional methods. Based on data-driven methods, the convolutional segmentation network can simultaneously produce pixel-level classification and location output [17]. The first segmentation-based SIRST detection method ACM was proposed by [18], which designed a semantic segmentation network using asymmetric context module. On this basis, Dai [19] further introduced expanded local contrast to improve their model. By combining traditional methods with deep learning methods and using bottom-up local attention modulation modules to embed subtle low-level details at higher levels, excellent detection performance was achieved. In [20], a balance between missed detection (MD) and false alarms (FA) was achieved by using cGAN networks to separately build models for MD and FA as two subtasks as generators. Next, a discriminator for image classification is used to distinguish the outputs of the two generators and ground-truth images. Zhang [21] uses attention mechanisms to guide the pyramid context network to detect targets. First, the feature map is partitioned to calculate local correlations. Second, the global contextual attention is used to calculate the correlation between semantics. Finally, the decoded images of different scales are fused to improve detection performance. Cheng [22] Using visible light imagery to achieve drone detection. First, the backbone network is lightly improved by using a multi-scale fusion method to improve the use of shallow features. To address the problem of drone loss in multi-scale detection, a novel non-maximum suppression method is developed to ultimately achieve real-time detection. However, the above methods still have many shortcomings. First, the problem of small target feature loss in the deep layers of the network still exists, and the contradiction between high-level semantic features and high-resolution cannot be resolved. Second, the coding maps generated by each downsampling layer cannot be well used. Overall, the above methods overlook the characteristics of drones. These problems will make the detection algorithm less robust to scene changes (such as cluttered backgrounds, targets with different SCR, shapes and sizes).

To solve these problems, we propose a data-driven progressive feature fusion detection method (PFFNet) from the perspective of infrared unmanned aerial vehicle target detection. First, global features were extracted from the input infrared image. Then, passes the encoding maps output by the downsampling to the FSM and PFM modules. The deep features that include high-level semantic information, the shallow features that contain rich image contour and the position information can be fully fused. Thereby improving the utilization of the output encoding maps of the downsampling layer. In addition, the output feature maps are cross-scale fused to enhance the response amplitude of infrared unmanned aerial vehicle targets in the deep network and solve the problem of feature loss in small targets in the deep layers of the network. The high-level semantic information and shallow semantic information are superimposed and output through dimensional cascading. The confidence map is obtained through threshold segmentation to output the final detection result. Finally, to verify the effectiveness of PFFNet, we conducted extensive ablation studies on FSM and PFM. And then conducted comparative experiments with existing methods on the SIRST Aug and IRSTD datasets. The experimental results show that the various modules of PFFNet have improved the detection of infrared unmanned aerial vehicle targets. Our algorithm has stronger robustness, better detection performance, and faster target detection time.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2023 doi:10.20944/preprints202305.0100.v1

2. Methods

Given an input image I, we aim to classify each pixel by end-to-end convolving a neural network to determine whether it is a drone target. Finally output a segmentation result that is the same size as I. The PFFNet detection algorithm is divided into two parts: the global feature extractor and the progressive feature fusion network. The global feature extractor extracts the basic features of the input infrared image I by looking at the entire image. The redundant information in the image can effectively reduce by obtaining these basic features.

The progressive fusion network is divided into two modules: the Neck and the Head. The Neck includes the Pool Pyramid Fusion Model (PFM) and the Feature Selection Model (FSM). The former is used to enhance the feature response amplitude in the deep network of the infrared drone target. The latter acts as a bridge for information interaction between high and low layers, increasing the utilization rate of the downsampling output encoding map. The Head implements the progressive fusion of feature maps of different scales and generates a segmentation mask.

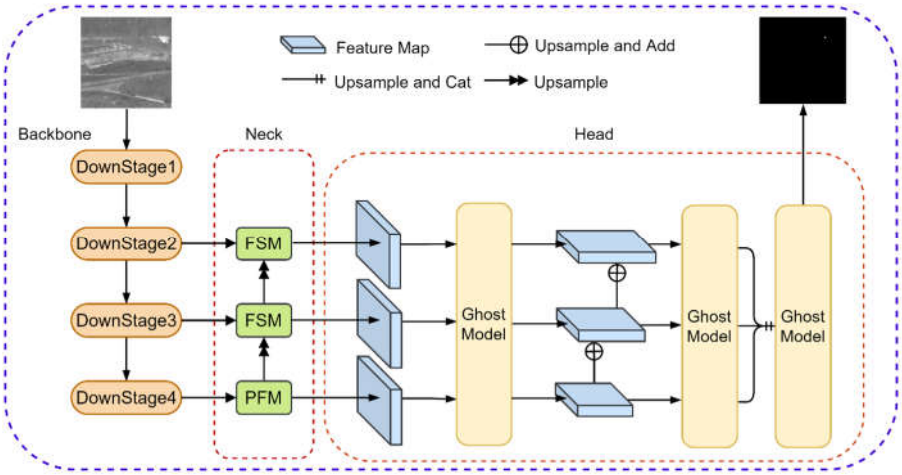


Figure 1. Structure of PFFNET

As shown in Figure 1, the input image I is encoded into different dimensions and resolutions by the backbone to generate encoding maps b_a (a=2, 3, 4). The low-level spatial position information of the target's salient features is obtained from b_a (a=2,3) by the FSM. Locating the high-frequency response area to reduce the influence of redundant signals on the target position information, and outputs the feature maps f_a (a=2,3). The b_4 is used as the input of the PFM to output the decoded image p. The PFM is composed of four different pooling structures in parallel to form a pyramid network. The high-frequency response amplitude of deep target features is enhanced and then passes it to the FSM after upsampling. The FSM and PFM extract local features of targets and use the progressive fusion method to calculate the phase output feature maps y_a (a=1,2,3). After being processed by the Ghost Model [23], y_a is doubled in size and element-wise added. This process greatly simplifies the task of small target detection by sharing the same weight for all convolution blocks, and reduces the parameters of the P algorithm by using element-wise addition while reducing the network inference time.

Then, the fused output is upsampled and dimensionally cascaded through convolution calculation. We proposed a multi-scale fusion strategy to progressively fuse feature maps of different sizes. Furthermore, the confidence map O is obtained by performing the final threshold segmentation on the fused feather map. Backbone is mainly used to expand the receptive field and extract deep semantic features. Upsampling helps to restore the size of the feature map. The progressive multi-scale feature fusion is achieved by upsampling and downsampling. The FSM and PFM modules are used to ensure the feature representation of small targets in the network.

To achieve good context data modeling ability, the simplest way is repeatedly and stack the network depth. The more layers the network has, the richer the semantic information and the larger

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2023 doi:10.20944/preprints202305.0100.v1

the receptive field [24-27]. However, infrared small targets have significant differences in size and a very low pixel ratio. If the network depth is blindly increased, the problem of feature disappearance may occur after the drone target undergoes multiple downsampling operations. Therefore, we should design special modules to extract high-level features while ensuring the representation of small targets with a very small pixel ratio in the deep network.

2.1. Feature Selection Module

The Feature Selection Module is mainly divided into two parts: LSM (Location Selection Model) and CSM (Channel Selection Model). Due to the small proportion in the image, drone targets are easy to lose or even weakening of the response amplitude of the target area during downsampling and upsampling. We found through experiments that there are rich target contour features in high-level semantic features, and accurate target location information in low-level semantic features. The FSM can use the semantic information of each dimension to achieve information interaction between different encoding maps. Through this module, the utilization rate of the downsampling and upsampling output encoding maps can be effectively increased. Besides, the effectiveness of multi-scale feature fusion can be guaranteed by locating and enhancing the high-frequency response amplitude area.

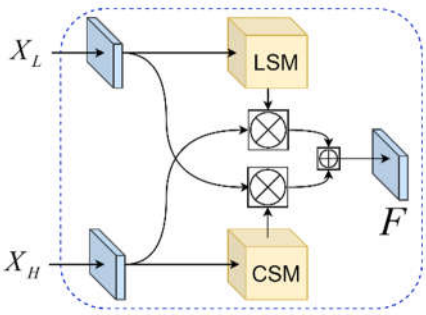


Figure 2. Structure of FSM

Figure 2 shows the feature representation of small targets retained in deep networks without losing spatial detail encoding of target positions. By combining LSM and CSM and utilizing CSM to enhance information interaction between high and low levels, obtaining target position through LSM. The combination forms the Feature Selection Module, which can fully integrate the deep features of high-level semantic information and the shallow features with rich image contour information and position information. And then improving the utilization rate of the encoding output map. The output of the feature selection module $F = \mathbb{R}^{C\times H\times W}$ can be represented as:

$$F = 2(C(X_H)\otimes X_L + L(X_L)\otimes X_H) \quad (1)$$

Where X_H is the deep feature that includes high-level semantic information, X_L is the shallow feature that contains rich image contour information and position information, ⊗ and ⊕ represents element-wise multiplication and addition of vectors, C and L represent the CSM and LSM modules, respectively.

2.2. Channel Selection Model

To solve the problem of losing or weakening the target area response value during upsampling of drone targets, we use CSM to enhance the target area response amplitude. As shown in Figure 3(a), the channel features at each spatial position are individually aggregated. The subtle details of deep drone targets are highlighted by directionally enhancing the high-frequency response channel weights of small targets. This module first performs average pooling and max pooling operations on the input feature map X to generate different 3D tensors x_i. Coupleing the global information of the feature map X in its internal channel. Then, a 1x1 convolution is used to evaluate the importance of

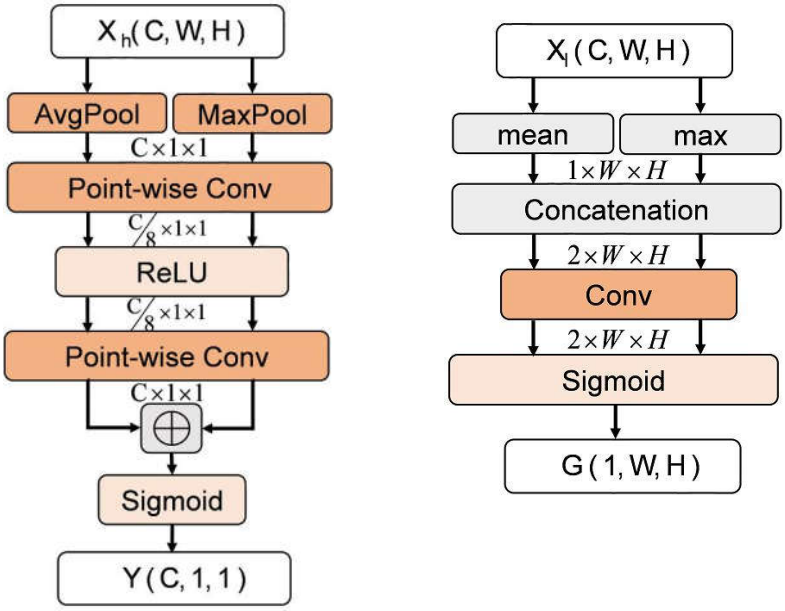
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2023 doi:10.20944/preprints202305.0100.v1

each channel and calculate the corresponding weight. The aggregated output $H(X) \in \mathbb{R}^{C\times H\times W}$ can be represented as:

$$H(X) = \sigma \sum_{i=1,2} P(X_i) = \sigma \sum_{i=1,2} \varepsilon_{1\times 1}\left(\delta\left(\varepsilon_{1\times 1}(x_i)\right)\right) \tag{2}$$

$$x_i = \frac{1}{H \times W} \sum_{i=j=1}^{H,W} X[:, i, j] \tag{3}$$

When i=1, x_1 is the feature vector obtained by average-pooling. When i=2, x_2 is the feature vector obtained by max-pooling. $\varepsilon_{1\times 1}$ are the point-wise convolutions with two convolution kernels of size 1×1 but different dimensions. δ represents the sigmoid function. σ represents the rectified linear unit, and output size of $(c, \frac{c}{r}, 1,1)$ and $(\frac{c}{r}, c, 1,1)$. Inspired by [28], this paper takes r=8 as the downsampling ratio for channel reduction.



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

Figure 3. Structure of CSM and LSM, (**a**) Channel selection model; (**b**) Location selection model.

2.3. Location Selection Model

The pixel number of infrared small targets in infrared images is extremely low, which lead to easily introduce interference signals during the process of feature extraction. LSM could be used to quickly locate local regions with visual saliency. As shown in Figure 3(b), this module calculates the maximum and mean values of the input feature map X and performs a cascade operation in the dimension direction. Then the module performs a convolution operation on the concatenated feature map. Here, a 7×7 convolution can further expand the receptive field of the convolution kernel, capturing areas with higher local response amplitudes from the lower-level network. In addition, the accurate position of the drone target in the entire feature map is calculated. The high response amplitude area $L(X) \in \mathbb{R}^{C\times H\times W}$ can be calculated using the following formula:

$$L(X) = \delta(\varepsilon_{7\times 7}(\mathbb{C}(x_1, x_2))) \tag{4}$$

$$x_i = M(X) \tag{5}$$

When i=1, M(*) takes the mean of the feature map X. When i=2, M(*) takes the maximum value of the feature map X; $\mathbb{C}$ represents the dimension cascade operation. The final output size of the feature map of this module is (1, W, H).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2023 doi:10.20944/preprints202305.0100.v1

2.4. PFM

Deeper neural networks can obtain more detailed semantic information of the target, but this method is not suitable for smaller targets. As the number of downsampling increases, the feature of drone targets (such as propellers and arms) weakens or even disappears. To solve this problem, this paper proposes a Pooling Pyramid Fusion Module (PFM) for infrared small target detection, which is used to process the encoding map of the highest downsampling layer. Due to the small target size, spatial dimension compression can be achieved through different global adaptive pooling layer structures. Besides, the corresponding dimension mean-value can be extracted to enhance the feature representation of small targets in deep networks. As shown in Figure 4, the input feature map $I \in \mathbb{R}^{C\times H\times W}$ is parallelly input into the pyramid pooling module for decoding, generating four encoding structures of different size 1×1, 2×2, 3×3, and 6×6. Then, 1×1 convolution is used to reduce the feature dimension to 1/4C. The four feature maps of different sizes are upsampled by bilinear interpolation. Then concatenating with the input feature map in the channel dimension. Finally, a 3×3 convolution is performed to output the feature map $O \in \mathbb{R}^{C\times H\times W}$, and form a contextual pyramid though five feature maps of the same dimension but different scales.

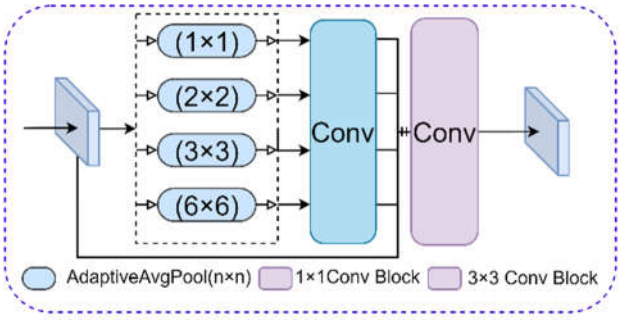


Figure 4. Structure of PFM

2.5. Segmentation Head

After multiple downsampling and convolution calculations, the targets' feature response in the deepest layer of the convolutional network will weaken. In response to this problem, we proposed a progressive feature fusion structure that is better suited for drones. As shown in Figure 5, this segmentation head can fuse different sizes of feature maps and enable the stacking of information between high and low layers to enhance the high-frequency response amplitude of the target. The input I with different sizes are proceed through Ghost Model, which is used to generate encoding maps with the same number and texture information through simple linear calculations. This reduces the convolution parameter volume and improves training and inference efficiency.

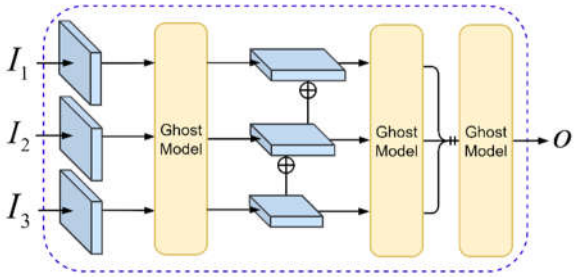


Figure 5. Structure of Head

PFFNet uses *SoftIoULoss* and *CELoss* to train the entire network and optimize the weighted loss between the predicted and segmented images. They can be expressed by the following formula:

$$CELoss = -\sum_{cls} T\log(P) \tag{6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2023 doi:10.20944/preprints202305.0100.v1

$$SoftIoULoss_{smooth} = \frac{\sum_{pixels} TP + smooth}{\sum_{pixels}(T + P - TP) + smooth} \tag{7}$$

where T and P represent the pixel values of the real target and the output prediction, respectively. Based on the initial loss value during training, α=3 and β=1 were set to balance the individual loss with the total loss to optimize the algorithm in the expected direction. To ensure the stability in the calculation, this paper sets *smooth*=1. Different weight balances may affect performance indicators [29].

3. Experiments

This section mainly introduces the implementation details and evaluation metrics of the algorithm, and compares it with other methods on two different datasets. In order to verify the effectiveness of the data-driven model PFFNet, comparative experiments and ablation experiments were conducted respectively.

3.1. Datasets

Performance based on data-driven methods is highly affected by the quality, quantity, and diversity of the data. Infrared image datasets have fewer images compared to visible datasets. Most methods are trained and evaluated on their own private datasets. Wang [20] established an open infrared small target datasets which includes 10k images. However, many of the targets in the dataset are too large and the annotations are not accurate which affects the training effectiveness. Dai [18] released a dataset with high-quality semantic segmentation masks. But the dataset is small, which easily leading to unstable model training, overfitting, and model convergence problems. To verify the reliability and robustness of the algorithm, experiments were conducted on two publicly available datasets with different image sizes (SIRST Aug [21] and IRSTD 1k [30]). Dataset [21] includes 8525 images in the training set and 545 images in the test set with an image size of 256×256, which is sufficient to satisfy the training requirements of data-driven models. The image size in dataset [30] is 512×512 and includes different types of small targets such as drones, organisms, ships, and vehicles. This dataset also covers many different scenes, including seawater, fields, mountains, cities, and clouds, with a cluttered background and severe noise, which is sufficient to verify small target detection methods.

3.2. Experimental Preparation and Evaluation Method

PFFNet will conduct ablation experiments and multi-algorithm comparison experiments on the SIRST Aug and IRSTD 1k open datasets.

We use classic semantic segmentation evaluation indicators as F1-score, receiver operating characteristic curve (ROC), and Intersection over Union (IoU). To measure the connection between precision and recall, *F1-score* is introduced. Meaning that the network must be able to detect targets and ensuring as few false alarms as possible. ROC is a qualitative indicator that reflects the connection between target detection rate (P_d) and false alarm rate (P_f). Precision, recall, target P_d, and P_f are defined as follows:

$$Precision = \frac{T_P}{T_P + F_P}, P_d = Recall = \frac{T_P}{T_P + F_N}, P_F = \frac{F_P}{N} \tag{8}$$

where T_P represents the target pixels that are correctly matched with the true label by the predicted pixels. F_P represents the background label pixels that are incorrectly predicted as targets. F_N represents the number of target pixels that are incorrectly classified as background. N represents the total number of pixels in the image. *F1-score* and IoU can be defined as:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}, \text{IoU} = \frac{T_P}{T + P - T_P} \tag{9}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2023 doi:10.20944/preprints202305.0100.v1

PFFNet is implemented based on Pytorch. The optimizer uses stochastic gradient descent (SGD), with momentum and weight decay coefficients set to 0.9 and 0.0001, respectively. The initial learning rate is 0.05, and a poly decay strategy is used. In SIRST Aug, the batch size is set to 32, and 30 epochs are trained. In IRSTD 1k, the batch size is set to 8, and 150 epochs are trained. In terms of hardware, we use a Tesla P100 GPU for training and a 3060 GPU for inference.

3.3. Comparative Experiments

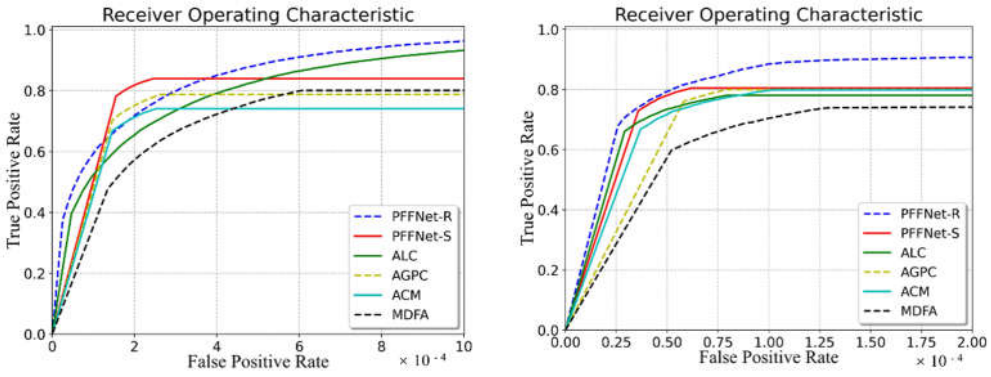
We compared PFFNet with four classic methods from different types. Results in [18][21] shown that data-driven methods are superior to model-driven methods. Therefore, only data-driven methods are compared in our experiments. In the data-driven scheme, we select AGPCNet, ACM, MDFA, ALC, PFFNet-R (PFFNet-ResNet-18[31]), and PFFNet-S (PFFNet-Swin Transformer v1[32]) for comparison. As shown in Table 1, the maximum and second values in each column are marked in bold and underlined, respectively. Obviously, PFFNet-R achieves the best results on both datasets, and its performance on the SIRST Aug dataset is better than on the IRSTD 1k dataset. This is because the IRSTD 1k dataset contains more challenging situations for detecting small infrared targets, including shape-changing targets and low contrast, as well as low SCR backgrounds with clutter and noise. Nevertheless, due to the effective aggregation of high-low level features and cross-layer features by the designed FSM and PFM modules, PFFNet still achieves the best results. In addition, to demonstrate the detection speed of the proposed algorithm for infrared small targets, we calculate the average running time of different methods on 1,000 infrared images (with a size of 256×256), where PFFNet is 6ms slower than ACM. But it achieves better detection results and can be used for real-time detection of drone targets.

Table 1. Results of the Comparation

Methods	SIRST Aug				IRSTD 1k				Time on GPU/s
	Precision	Recall	IoU	F1-score	Precision	Recall	IoU	F1-score	
ACM	87.24	70.72	64.09	78.12	76.57	74.79	60.86	75.67	**0.005**
ALC	<u>88.00</u>	74.94	71.86	80.95	<u>80.25</u>	73.40	62.16	76.67	0.058
MDFA	81.05	65.38	56.71	72.38	66.52	69.96	51.74	68.20	0.064
AGPCNet	87.73	74.71	67.64	80.70	76.06	<u>77.98</u>	62.62	77.01	0.052
PFFNet-S	81.04	**88.99**	<u>73.66</u>	<u>84.83</u>	78.44	**77.87**	<u>64.15</u>	<u>78.16</u>	<u>0.011</u>
PFFNet-R	**88.74**	<u>81.28</u>	**73.68**	**84.85**	**81.84**	77.40	**66.05**	**79.56**	0.023

Furthermore, to visually compare the AUC, the ROC curves of these methods on two different datasets are shown in figure 6. These experimental results demonstrate that PFFNet can greatly suppress the background and fully leverage the algorithm's advantages. This method fully learns highly discriminative semantic features from diverse training data to achieve highly robust object detection results. In addition, it can segment targets more accurately than other state-of-the-art methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2023 doi:10.20944/preprints202305.0100.v1



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

Figure 6. ROC curve, (**a**) SIRST Aug dataset; (**b**) IRSTD 1k dataset.

The analysis of the experimental results is presented in figure7, which shows the mask information of the target and its 3D display in four different representative scenes. Image a and d have multiple drone targets but the local contrast is low between target and background. Image b and c have higher local contrast but the background is relatively complex.

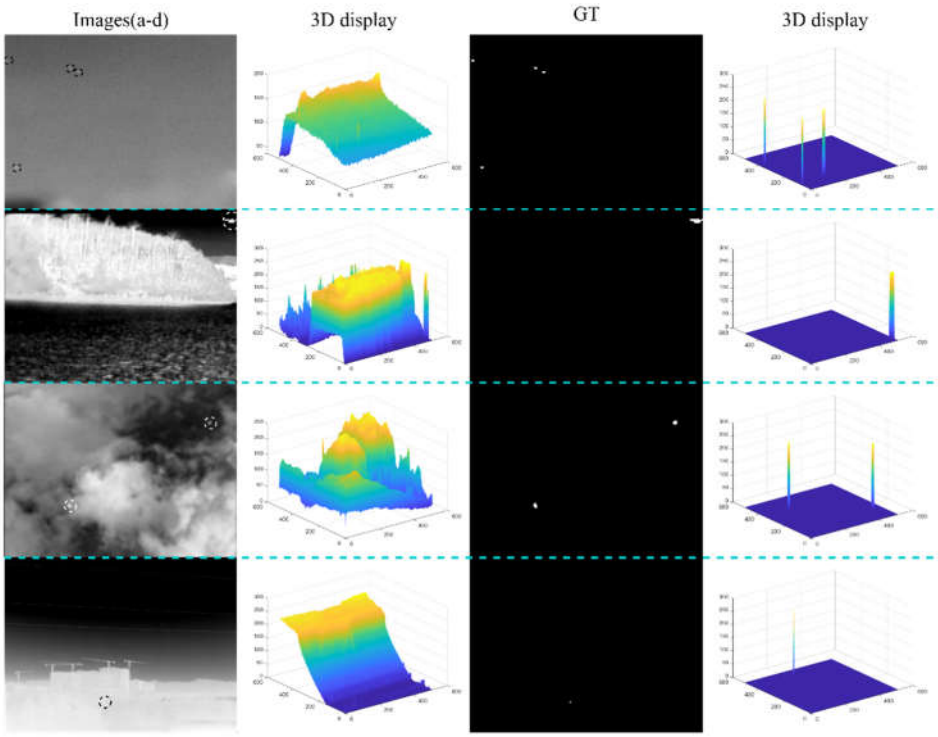


Figure 7. The performance of the algorithm in 4 different images.

This article performs target detection on these four different scenes, and the output 3D displays of six different algorithms are shown in the figure8. There is a lot of noise similar to the target in Fig7. a and Fig7. d. ACM and MDFA cannot make correct predictions, but PFFNet can clearly distinguish between the target and the background. This article's model also performs well on multiple targets in Fig7. a. In contrast, ALC cannot distinguish all targets. The overall performance of ALC is good, but it may have some false alarms in situations with high noise levels or low SCR. PFFNet-S performs similarly to PFFNet-R, where the former can detect the target contour information and the latter can locate the target's specific position. Results proves that PFFNet has the best detection results in various scenes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2023 doi:10.20944/preprints202305.0100.v1

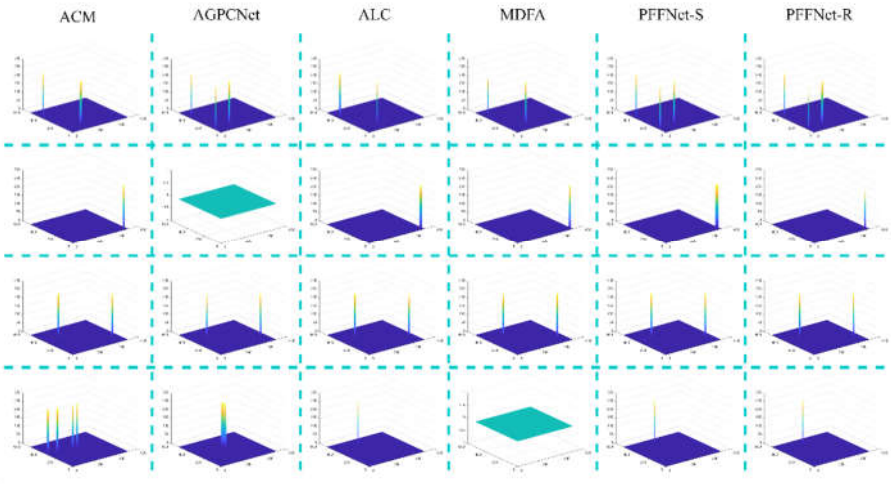


Figure 8. 3d display of algorithm output

3.4. Ablation Experiments

To verify the detection performance and speed of PFFNet, this paper conducted ablation experiments with different settings on the SIRST Aug dataset. The effectiveness of CPM, AFM, and Segmentation Head was verified by controlling their addition in Swin Transformer-v1 [32]. To prevent overfitting caused by the excessive parameterization, the last layer of Swin Transformer-v1 was removed. With only three downsampling layers replaced. Meanwhile, the pre-trained weights were used. Since FSM and PFM are based on the Head, the backbone and head were selected as the basic network for ablation experiments. In addition, ResNet-18 [31] was also selected as the backbone to verify the portability of PFFNet. As shown in Table 2, under the same basic network, the performance on the dataset was improved by adding FSM, PFM, and Head. Moreover, PFFNet still achieved good results with different feature extraction networks.

Table 2. Results of the ablation

Backbone	Head	Neck		SIRST Aug	
		FSM	PFM	IoU	F1-score
Swin T				47.78	55.02
Swin T	√			71.26	82.44
Swin T	√	√		73.42	84.69
Swin T	√		√	71.63	83.08
Swin T	√	√	√	73.66	84.83
ResNet-18				68.12	74.42
ResNet-18	√	√	√	**73.68**	**84.85**

In FSM, different experiment were proceeded to verify the effectiveness of CSM and LSM, as shown in Table 3. Swin Transformer-v1 and head were used as the basic network, and CSM and LSM modules were added separately. Both of them show good performance improvement, especially LSM. By enhancing the response weight of local regions with visual saliency, the position information of the target can be obtained to improve the overall performance of the network.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2023 doi:10.20944/preprints202305.0100.v1

Table 3. Validation of the modules

Backbone	CSM	LSM	SIRST Aug			
			IoU	F1-score	Precison	Recall
Swin T + head			71.26	82.44	77.10	88.57
Swin T + head	√		72.46	83.62	**78.59**	90.28
Swin T + head		√	**72.81**	**84.26**	77.98	**91.65**

Dimensionality reduction can reduce redundant information and greatly accelerate the training speed in network. But this may loss of useful information. This paper selects different dimensionality reduction ratios to explore the best way to segment small infrared targets. In PFFNet, two dimensionality reductions were performed in FSM and PFM separately, and the reduction ratios were denoted as (r_f, r_p). According to [28], we set r_f = 8. As shown in Table 4, the best result was obtained when (r_f = 8, r_p = 4). Although the performance on the dataset may vary with different parameter settings, the overall change is not significant.

Table 4. Ratio of dimension reduction

Reduction Ratios	SIRST Aug			
	IoU	F1-score	Precision	Recall
(8,1)	71.8	83.58	77.80	90.29
(8,2)	73.05	84.43	80.08	89.27
(8,4)	**73.66**	**84.83**	**81.04**	88.99
(8,8)	70.62	82.78	74.77	**92.71**

Table 5 shows the inference time of each module based on two different feature extractors, using 1000 infrared images with a size of 256×256 as the benchmark and taking the average. In which PFFNet-S has the shortest inference time. There is a gap in the total inference time between the two algorithms due to the difference backbone, but the gap between each submodule is very small. It can be seen from Table 5 that PFFNet has a good real-time performance. In general, experiments show that PFFNet is a lightweight network that can quickly and accurately detect infrared drone targets.

Table 5. Time of module inference

Methods	Backbone	PFM	FSM	Head	PFFNet	ALL
PFFNet-S	7.40	0.78	1.69	1.48	3.95	11.24
PFFNet-R	12.50	1.41	3.32	6.31	11.04	23.56

4. Conclusions

This paper proposes a fast detection method for infrared small targets: Progressive Feature Fusion Network (PFFNet). Faced the problem of losing target area response values during downsampling of drone targets, FSM is proposed. It can fully fuse deep features with high-level semantic information and shallow features with rich image contour and location information. Successfully achieved information exchange between downsampling layers. Then, PFM is proposed to integrate deep features and enhance high-frequency response amplitude from a multi-scale perspective to address the problem of weakened small target feature representation in deep networks. Meanwhile, a lightweight segmentation head suitable for infrared small targets is designed to progressively fuse low-level and high-level semantics from the perspective of feature fusion. In addition, the utilization of features in downsampling layers was improved. Finally, module comparison is conducted on two datasets with different complexities, which fully demonstrates the effectiveness of each module. The practicality verification and inference time statistics on the SIRST Aug dataset confirm that PFFNet has a good performance for fast detection of infrared small targets. As well, a large number of data-driven algorithm comparison experiments demonstrate the ability of

PFFNet to cope with complex scene detection tasks in term of numerical evaluation. Meanwhile, PFFNet has better detection performance and shorter inference time.

However, there are still some problems of the algorithm that need further research. Such as dealing with network overfitting, utilizing more efficient contextual information, etc. In future work, attention mechanisms and fusion structures will continue to be explored for their application in infrared drone target detection.

Author Contributions: Conceptualization, C.Z. and Z.H.; methodology, Z.H.; software, Z.H.; validation, Z.H., K.Q. and M.Y.; formal analysis, M.Y.; investigation, H.F.; resources, C.Z.; data curation, C.Z.; writing—original draft preparation, Z.H.; writing—review and editing, C.Z.; visualization, Z.H.; supervision, C.Z.; project administration, C.Z.; funding acquisition, C.Z. and M.Y.

Funding: This research was funded by Liaoning Provincial Department of Education, grant number LJKMZ20220605.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. Original data can be obtained by contacting the corresponding author.

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

References

1. Kapoulas, I.K.; Hatziefremidis, A.; Baldoukas, A.K.; Valamontes, E.S.; Statharas, J.C. Small Fixed-Wing UAV Radar Cross-Section Signature Investigation and Detection and Classification of Distance Estimation Using Realistic Parameters of a Commercial Anti-Drone System. *Drones* **2023**, *7*, 39. https://doi.org/10.3390/drones7010039.
2. Wang, C.; Meng, L.; Gao, Q.; Wang, J.; Wang, T.; Liu, X.; Du, F.; Wang, L.; Wang, E. A Lightweight Uav Swarm Detection Method Integrated Attention Mechanism. *Drones* **2022**, *7* (1), 13. https://doi.org/10.3390/drones7010013.
3. Bai X, Zhou F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition*. 2010;43(6):2145-2156. doi:10.1016/j.patcog.2009.12.023.
4. Chang B, Meng L, Haber E, Ruthotto L, Begert D, Holtham E. Reversible Architectures for Arbitrarily Deep Residual Neural Networks. Published online November 18, 2017. doi:10.48550/arXiv.1709.03698.
5. Rivest JF, Fortin R. Detection of dim targets in digital infrared imagery by morphological image processing. *OE*. 1996;35(7):1886-1893. doi:10.1117/1.600620.
6. Chen CLP, Li H, Wei Y, Xia T, Tang YY. A Local Contrast Method for Small Infrared Target Detection. IEEE Transactions on Geoscience and Remote Sensing. 2014;52(1):574-581. doi:10.1109/TGRS.2013.2242477.
7. Wei Y, You X, Li H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognition*. 2016;58:216-226. doi:10.1016/j.patcog.2016.04.002.
8. Han J, Ma Y, Zhou B, Fan F, Liang K, Fang Y. A Robust Infrared Small Target Detection Algorithm Based on Human Visual System. *IEEE Geoscience and Remote Sensing Letters*. 2014;11(12):2168-2172. doi:10.1109/LGRS.2014.2323236.
9. Han J, Moradi S, Faramarzi I, Liu C, Zhang H, Zhao Q. A Local Contrast Method for Infrared Small-Target Detection Utilizing a Tri-Layer Window. *IEEE Geoscience and Remote Sensing Letters*. 2020;17(10):1822-1826. doi:10.1109/LGRS.2019.2954578.
10. Zhang L, Peng Z. Infrared Small Target Detection Based on Partial Sum of the Tensor Nuclear Norm. *Remote Sensing*. 2019;11(4):382. doi:10.3390/rs11040382.
11. Zhu H, Liu S, Deng L, Li Y, Xiao F. Infrared Small Target Detection via Low-Rank Tensor Completion With Top-Hat Regularization. *IEEE Transactions on Geoscience and Remote Sensing*. 2020;58(2):1004-1016. doi:10.1109/TGRS.2019.2942384.
12. Dai Y, Wu Y, Song Y, Guo J. Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values. *Infrared Physics & Technology*. 2017;81:182-194. doi:10.1016/j.infrared.2017.01.009.
13. Fu J, Liu J, Tian H, et al. Dual Attention Network for Scene Segmentation. Published online April 21, 2019. doi:10.48550/arXiv.1809.02983.
14. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *In Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, USA, 7–12 February 2020; pp. 12993–13000. doi:10.1609/aaai.v34i07.6999.
15. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Cheng, Y.; Alexander, C.B. SSD: Single Shot MultiBox Detector. In European Conference on Computer Vision; Springer: Cham, Switzerland, 2016; pp. 21–37. doi:10.1007/978-3-319-46448-0_2.

16. Wang, C.; Shi, Z.; Meng, L.; Wang, J.; Wang, T.; Gao, Q.; Wang, E. Anti-Occlusion UAV Tracking Algorithm with a Low-Altitude Complex Background by Integrating Attention Mechanism. *Drones* 2022, 6, 149.
17. Liu, S.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. *In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
18. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric Contextual Modulation for Infrared Small Target Detection.; 2021; pp. 950–959. Doi: https://arxiv.org/abs/2009.14530v1.
19. Dai Y, Wu Y, Zhou F, Barnard K. Attentional Local Contrast Networks for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*. 2021;59(11):9813-9824. doi:10.1109/TGRS.2020.3044958.
20. Huan Wang, Luping Zhou, and Lei Wang. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 85098518, 2019. 2, 3, 6
21. T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-Guided Pyramid Context Networks for Detecting Infrared Small Target Under Complex Background," pp. 1–13, 2023, doi: 10.1109/TAES.2023.3238703.
22. Cheng, Q.; Wang, H.; Zhu, B.; Shi, Y.; Xie, B. A Real-Time UAV Target Detection Algorithm Based on Edge Computing. *Drones* **2023**, 7, 95, doi:10.3390/drones7020095.
23. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. arXiv March 13, 2020. http://arxiv.org/abs/1911.11907.
24. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
25. 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
26. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In European Conference on Computer Vision; Springer: Cham, Switzerland, 2018; pp. 122–138.
27. Xiong, Y.; Liu, H.; Gupta, S.; Akin, B.; Bender, G.; Wang, Y.; Kindermans, P.J.; Tan, M.; Singh, V.; Chen, B. MobileDets: Searching for Object Detection Architectures for Mobile Accelerators. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3824–3833.
28. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I. S. CBAM: Convolutional Block Attention Module. arXiv July 18, 2018. http://arxiv.org/abs/1807.06521.
29. Chen, Y.; Li, L.; Liu, X.; Su, X. A Multi-Task Framework for Infrared Small Target Detection and Segmentation. *IEEE Trans. Geosci. Remote Sensing* **2022**, *60*, 1–9. https://doi.org/10.1109/TGRS.2022.3195740.
30. M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape Matters for Infrared Small Target Detection," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 867–876. doi: 10.1109/CVPR52688.2022.00095.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. arXiv December 10, 2015. http://arxiv.org/abs/1512.03385.
32. Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Published online August 17, 2021. Accessed February 20, 2023. http://arxiv.org/abs/2103.14030.