drones

MDPI

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 110159, China; hanzq@sylu.edu.cn (Z.H.); ymkig@sylu.edu.cn (M.Y.); quankn@stu.sylu.edu.cn (K.Q.)
[2] Science and Technology on Electromechanical Dynamic Control Laboratory, Beijing Institute of Technology, Beijing 100081, China; 7520210110@bit.edu.cn
* Correspondence: zcbxl@sylu.edu.cn

Abstract: The rampant misuse of drones poses a serious threat to national security and human life. Currently, CNN (Convolutional Neural Networks) are widely used to detect drones. However, small drone targets often reduced amplitude or even lost features in infrared images which traditional CNN cannot overcome. This paper proposes a Progressive Feature Fusion Network (PFFNET) and designs a Pooling Pyramid Fusion (PFM) to provide more effective global contextual priors for the highest downsampling output. Then, the Feature Selection Model (FSM) is designed to improve the use of the output coding graph and enhance the feature representation of the target in the network. Finally, a lightweight segmentation head is designed to achieve progressive feature fusion with multi-layer outputs. Experimental results show that the proposed algorithm has good real-time performance and high accuracy in drone target detection. On the public dataset, the intersection over union (IOU) is improved by 2.5% and the detection time is reduced by 81%.

Keywords: background clutter; counter-drones; progressive fusion; lightweight network

Citation: Han, Z.; Zhang, C.; Feng, H.; Yue, M.; Quan, K. PFFNET: A Fast Progressive Feature Fusion Network for Detecting Drones in Infrared Images. *Drones* **2023**, *7*, 424. https://doi.org/10.3390/drones7070424

Academic Editor: Giordano Teza

Received: 1 May 2023
Revised: 19 June 2023
Accepted: 22 June 2023
Published: 26 June 2023

1. Introduction

Various types of small unmanned aerial vehicles threaten infrastructure, hardware and people seriously [1,2]. 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 flight altitude being usually below 500 m, drone targets often have a few to several tens of pixels in an infrared image. In addition, drones usually have a low signal-to-noise ratio (SCR) and are easily submerged in strong noise and cluttered backgrounds [3–5]. Therefore, the radiation intensity of the target is lower and it lacks significant morphological features, making target detection in infrared images difficult [6,7].
(2) The contradiction between the target and the detection algorithm: compared to RGB images, detecting drones in infrared images presents more problems, such as the lack of shape and texture features. After filtering and convolution calculations, it is easy to weaken or even lose the representative features of drones (such as wings) [8,9]. 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 [10].

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 [11]. In addition, most negative samples are easily classified, which makes it difficult for the algorithm to

optimize in the expected direction [12]. Therefore, the nets designed for normal objects are hardly used to detect drones in infrared images.

To detect drone targets in infrared images, researchers have proposed many traditional methods over the past few years. The traditional detection method involves implementing SIRST (Single-frame InfraRed Small Target) detection by calculating the non-coherence between the target and background. Typical methods include filter-based methods [13–15], which can only suppress uniform and smooth background noise. They are also unable to adapt to complex backgrounds and may have a higher false alarm rate. The HVS method [16–19] uses the ratio of gray values between each pixel position and its neighboring region as an enhancement factor. It can effectively enhance the real target. However, it cannot effectively suppress the background noise. The low-rank representation [20–22] 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 calculations without training or learning. However, designing hand-crafted features and tuning hyperparameters require expert knowledge and a significant amount of engineering effort.

With the development of CNN methods, more data-driven methods are being applied to infrared small target detection [23–26]. 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 [27]. The first segmentation-based SIRST detection method ACM was proposed by [28], which designed a semantic segmentation network using an asymmetric context module. On this basis, Dai [29] further introduced expanded local contrast to improve their model. They use bottom-up local attention modulation modules to embed subtle low-level details at higher levels by combining traditional methods with deep learning methods. Zhang [30] used an attention mechanism to guide the pyramid context network to detect targets. The feature map is divided into blocks to compute local correlations. Then global contextual attention is used to compute the correlation between semantics. Finally, the decoding maps of different scales are fused to improve detection performance.

As well as using segmentation networks to solve the problem of drone target detection, researchers also offer some other ideas. In [31], a balance between missed detection (MD) and false alarms (FA) was achieved. The cGAN networks were used to separately build models for miss detection (MD) and false alarm (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. Chen [32] uses visible light images to achieve drone detection. By lightweight improvement of the backbone network and using the multi-scale fusion method to enhance the use of shallow features. A new non-maximum suppression method is designed to solve the problem of drone loss in multi-scale detection, ultimately achieving 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. These problems will make the drone detection algorithm less robust to scene changes (such as cluttered backgrounds, and 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, it passes the encoding maps output by 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 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.

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, we 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.

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 output 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 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 [33], 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. The 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 to repeat and stack the network depth. The more layers the network has, the richer the semantic information and the larger the receptive field [34–37]. 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.

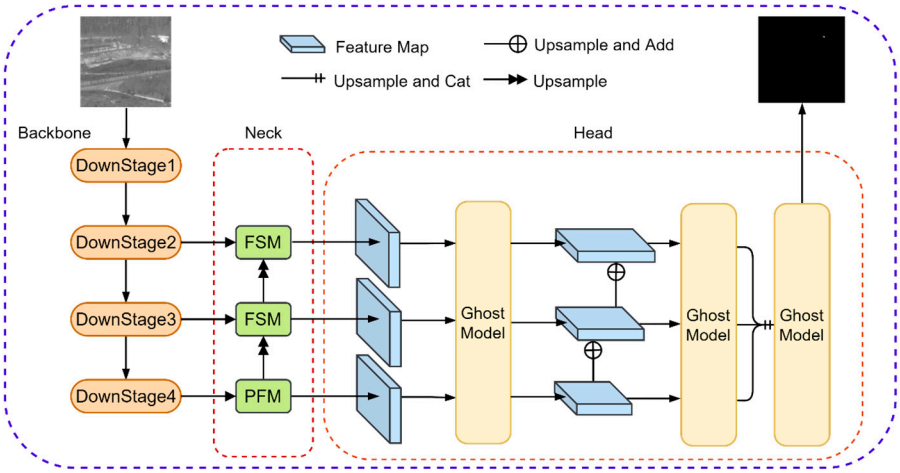


Figure 1. Structure and composition of PFFNET.

2.1. Feature Selection Module

We hope to improve the ability of channel information to interact between different downsampling layers based on a feature fusion perspective. Inspired by CBAM [38] and ShuffleNet [35], a new module called Feature Selection Module (FSM) is proposed to aggregate low-level and deep semantic information. The FSM is mainly divided into two parts: Location Selection Model (LSM) and Channel Selection Model (CSM). Experiments have shown that high-level semantic features contain detailed semantic information about the target, while low-level semantic features contain precise information about the target's location. As shown in Figure 2, FSM processes the different categories of information separately and uses the semantic information of each dimension to achieve information interaction between different coding graphs.

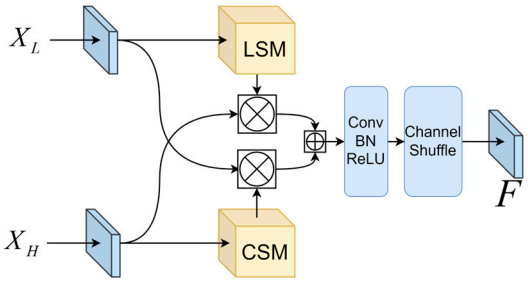


Figure 2. Structure and composition of Feature Selection Module.

To preserve the feature of drone targets in deep networks and encode the spatial details of targets' positions, this paper combines LSM and CSM. In order to cover more parts of the target to be recognized with features, this module first aggregates the channels via CSM. LSM was used to obtain more contour features and accurate location information. Then, the output of the current branch is multiplied by the input of another branch to enhance the high-low level channel features and facilitate multi-channel information interaction. Through 5×5 convolution we obtain a larger receptive field. Finally, the information exchange between different channels is established based on the Channel Shuffle idea. The output of the feature selection module $F = \mathbb{R}^{C \times H \times W}$ is represented as:

$$F\prime = \mu(C(X_H) \otimes X_L \oplus L(X_L) \otimes X_H) \quad (1)$$

$$F = \sigma(\varepsilon_{5\times5}(F\prime)) \tag{2}$$

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, $\otimes$ and $\oplus$ represents element-wise multiplication and addition of vectors, C and L represent the CSM and LSM modules, respectively. E is the convolution calculation. Σ represents the activation function of the Rectified Linear Unit. M is used to enhance feather representation and is a positive integer.

2.2. Channel Selection Model

To solve the problem of losing or weakening the target area response value during up-sampling of drone targets, the CSM is used to enhance the target area response amplitude.

As shown in Figure 3a, the channel features at each spatial position are individually aggregated. The high-frequency response channel weights of small targets are directionally enhanced to highlight the subtle details of deep targets. In Equation (3), this module performs average pooling and max pooling operations on the input feature map X_H to generate different 3D tensors x_{hi}. Coupling the global information of the feature map X in its internal channel. Then, a 1×1 convolution is used to evaluate the importance of each channel and calculate the corresponding weight. The aggregated output $H(X) \in \mathbb{R}^{C\times H\times W}$ can be represented as:

$$x_{h1} = \frac{1}{H \times W}\sum_{w=h=1}^{H,W} X_H[:, w, h] \tag{3}$$

$$x_{h2} = \max(X_H[:, W, H]) \tag{4}$$

where x_{h1} and x_{h2} are the feature vector calculated by average pooling and maximum pooling. W and h represent the width and height of the feature map, respectively. The output of CSM as $C(X_H) \in \mathbb{R}^{C\times H\times W}$ is

$$C(X_H) = \sigma \sum_{i=1,2} \varepsilon_{1\times1}(\delta(\varepsilon_{1\times1}(x_{hi}))) \tag{5}$$

where δ represent the Sigmoid function. The output through $\varepsilon_{1\times1}$ are (c, c/r_f, 1, 1) and (c/r_f, c, 1, 1), respectively. R_f is the channel descent ratio.

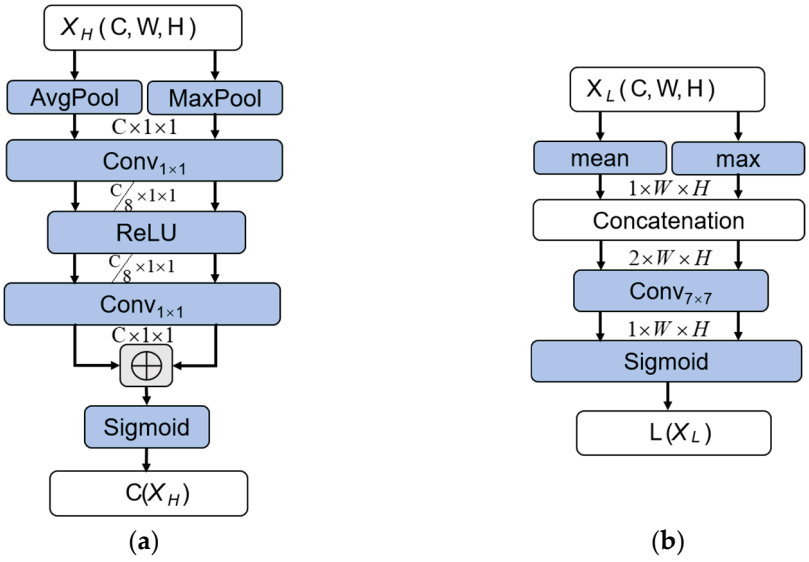


Figure 3. Structure and composition of CSM and LSM, (**a**) Channel selection model; (**b**) Location selection model.

2.3. Location Selection Model

The SCR of drone targets in infrared images is extremely low, which easily take interference signals into the process of feature extraction. The LSM could quickly locate visual salient regions. As shown in Figure 3b, this module calculates the maximum and mean values of the input feature map X_L, respectively.

$$x_{l1} = \frac{1}{C}\sum_{1}^{C} X_L[c,:,;] \tag{6}$$

$$x_{l2} = \max(X_L[c,:,;]) \tag{7}$$

where x_{l1} and x_{l2} represent the mean and maximum calculation of channel dimension. Perform cascading operations in the channel direction before performing convolution operations. Here, a 7×7 convolution can further expand the receptive field of the convolution kernel. It can also capture areas with higher local response amplitudes from the lower-level network. In addition, the accurate position of the drone target in the feature map is ensured. The output $L(X) \in \mathbb{R}^{C\times H\times W}$ can be calculated using the following equation:

$$L(X_L) = \delta(\varepsilon_{7\times 7}(\mathbb{C}(x_{l1}, x_{l2}))) \tag{8}$$

where C represents the dimension cascade operation. The final output size of the feature map of this module is (1, W, H).

2.4. Pooling Pyramid Fusion Module

Deeper neural networks can obtain more detailed semantic information about the target, but this method is not suitable for smaller targets. As the number of downsamplings 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 drone target detection. Affected by [39], the PFM is only used to process the encoding map of the highest downsampling layer. Therefore, it can provide a more effective global context prior to pixel-level scene parsing. The PFM compresses spatial dimensions 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 sizes 1×1, 2×2, 3×3, and 6×6. Then, 1×1 convolution is used to reduce the feature dimension to 1/4C. r_p is the channel descent ratio. 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 through five feature maps of the same dimension but different scales.

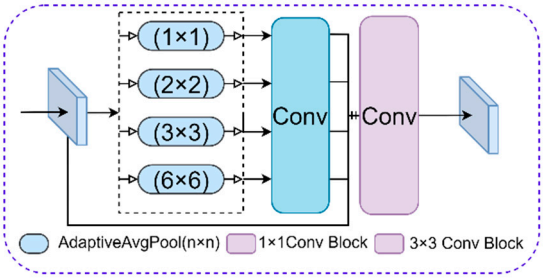


Figure 4. Structure and composition of PFM.

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. This segmentation head is designed and improved based on FPN [40]. It 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.

As shown in Figure 5, the input I with different sizes proceeds through the Ghost Model. Due to the small proportion of drone targets in the image, ordinary convolution calculations can generate a large number of feature maps with the same texture information or even without drone target information. However, the Ghost Model could generate encoding maps with the same number and texture information through simple linear calculations. This step reduces the convolution parameter volume and improves training and inference efficiency.

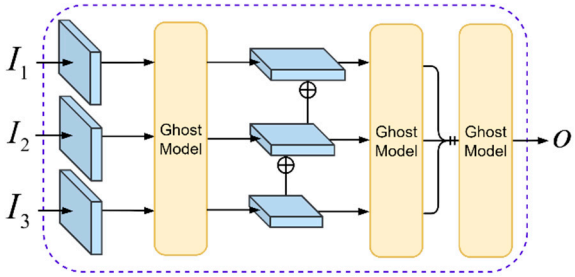


Figure 5. Structure and composition of segmentation head.

Finally, we train the entire network by *SoftIoULoss* and *CELoss* in PFFNet and optimize the weighted loss between the predicted and segmented images. They can be expressed by the following equations:

$$CELoss = -\sum_{cls} T\log(P) \tag{9}$$

$$SoftIoULoss_{smooth} = \frac{\sum_{pixels} TP + smooth}{\sum_{pixels} (T + P - TP) + smooth} \tag{10}$$

$$Loss = \alpha(1 - IOU) + \beta(1 - CELoss) \tag{11}$$

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 stability in the calculation, this paper sets *smooth* = 1. Different weight balances may affect performance indicators [10].

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 [31] established an open infrared small target dataset that includes 10,000 images. However, many of the targets in the dataset are too large and the annotations are not accurate

which affects the training effectiveness. Dai [28] released a dataset with high-quality semantic segmentation masks. However, the dataset is small, which easily leads 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 [30] and IRSTD 1k [41]). The dataset [30] 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 the dataset [41] 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. Therefore, it is sufficient to verify the proposed detection method.

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.

This paper uses classic semantic segmentation evaluation indicators such 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 ensure as few false alarms as possible. ROC is a qualitative indicator that reflects the connection between the target detection rate (P_d) and the false alarm rate (P_f). *Precision*, *recall*, target P_d, and P_f are defined as follows:

$$Precision = \frac{T_P}{T_P + F_P}, P_d = Recall = \frac{T_P}{T_P + F_N}, P_F = \frac{F_P}{N} \tag{12}$$

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:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}, \text{IoU} = \frac{T_P}{T + P - T_P} \tag{13}$$

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 of different types. Results in [28,30] show 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 and PFFNet for comparison. PFFNet-S and PFFNet-R represent the selection of Swin Transformer V1 [42] and ResNet-18 [43] as the global feature extractors, respectively. The purpose is to verify whether the algorithm is compatible with different feature extraction structures. The hyperparameters of other algorithms are not changed and remain at their default settings.

As shown in Table 1, the maximum and secondary values of each column are highlighted in bold and underlined, respectively. PFFNet achieved the best detection performance on both different datasets, with an IoU of up to 73.7% on the SIRST AUG dataset. PFFNet-R and AGPCNet were selected for comparison because they both chose ResNet-18 as their global feature extractor. Obviously, compared to AGPCNet, PFFNet-R increased IoU by 8.9% (73.7 vs. 67.6) and 5.5% (66.1 vs. 62.6) on both datasets, respectively. This method reduces detection time by 55.8% (0.01 vs. 0.05) while maintaining detection accu-

racy. Compared to ALC, PFFNet-S increased IoU by 2.5% (73.7 vs. 71.9) and 3.2% (64.2 vs. 62.2), respectively, and reduced the detection time by 81% (0.01 vs. 0.06). Furthermore, we calculated the average time of different methods on 1000 infrared images, among which PFFNet-S is 6 ms slower than ACM, but PFFNet-S has a better detection performance.

Table 1. Comparisons with other methods on SIRST Aug and IRSTD-1k.

Methods	SIRST Aug (%)				IRSTD 1k (%)				Time on GPU/s
	Precision	Recall	IoU	F1-Score	Precision	Recall	IoU	F1-Score	
ACM	87.2	70.7	64.1	78.1	76.6	74.8	60.9	75.7	**0.005**
ALC	88	74.9	71.9	81	80.3	73.4	62.2	76.7	0.058
MDFA	81.1	65.4	56.7	72.4	66.5	70	51.7	68.2	0.064
AGPCNet	87.7	74.7	67.6	80.7	76.1	78	62.6	77	0.052
PFFNet-S	81	**89**	73.7	84.8	78.4	**77.9**	64.2	78.2	0.011
PFFNet-R	**88.7**	81.3	**73.7**	**84.9**	**81.8**	77.4	**66.1**	**79.6**	0.023

Besides, it can also be seen that each algorithm performs significantly better on the SIRST Aug dataset compared to IRSTD 1k. Just like the IoU values of PFFNet-S on different datasets (73.7% on SIRST Aug dataset and 64.2% on IRSTD 1k dataset). There are two reasons for this result:

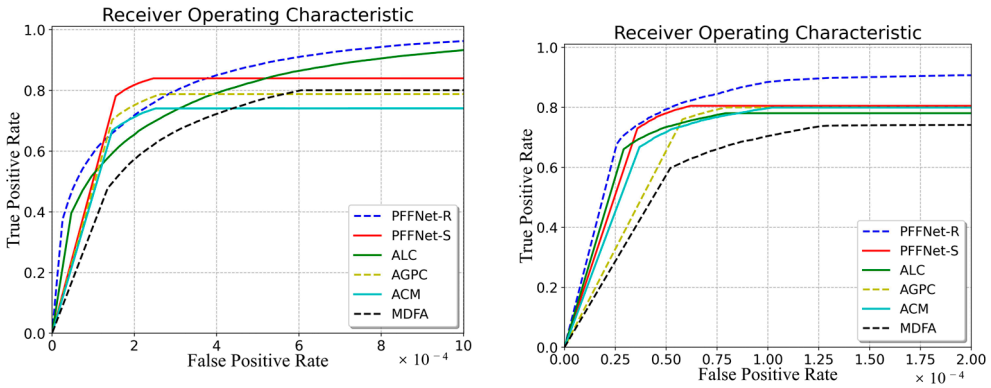
(1) Number of data used for model training

The results of detection based on data-driven methods are affected by the number of data. The SIRST Aug dataset is used to train models with more data than IRSTD 1k (8525:800). Due to the higher resolution of the images contained in IRSTD 1k (512 × 512), good detection results were also achieved even with a small amount of data. The higher the image resolution, the better the detection result, which has been proven [44].

(2) The complexity of the data

The IRSTD 1k dataset contains more challenging scenarios for detection. For example, drone targets with different shapes, low contrast, low signal-to-noise ratio, as well as more complex backgrounds and more noise interference. The complexity of its data is much higher than SIRST Aug.

Furthermore, to visually compare the AUC, the ROC curves of these methods on two different datasets are shown in Figure 6. Experimental results demonstrate that PFFNet can greatly suppress the background. 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.



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

Figure 6. The ROC curve compared with other methods on different data sets, (**a**) SIRST Aug dataset; (**b**) IRSTD 1k dataset.

This paper also selects four types of data for visual evaluation of the above algorithms. As shown in Figure 7, Introducing infrared images and GT (Ground Truth) for four different scenarios. The 'ori' represents the original images and the 'after' represents the enhanced ones. The local contrast between the target and background is extremely low in Figure 7a,d. Figure 7b,c have a high local contrast with a complex background. Meanwhile, Figure 7a,c have multiple drone targets.

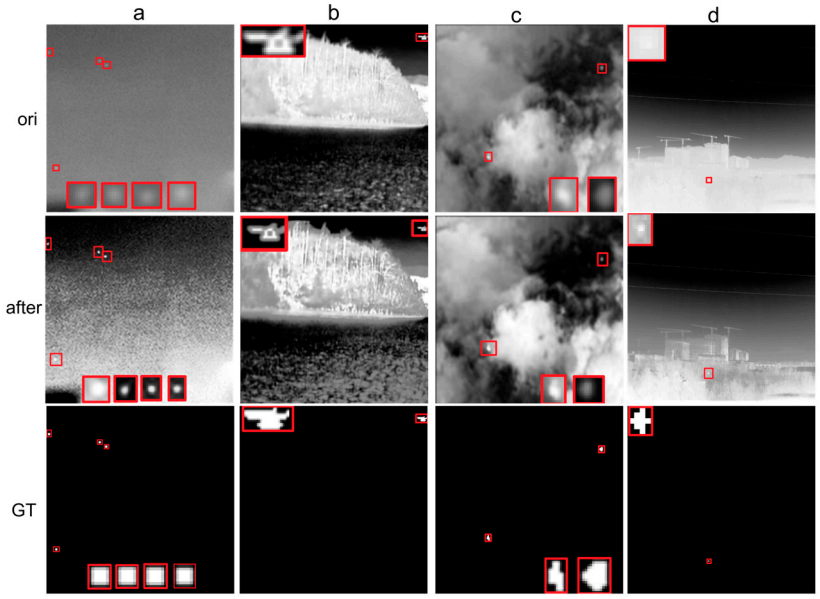


Figure 7. Four scenarios with drones and their GT.

Figure 8 shows the output of the GT map with five different methods. Although the enhanced after-a has stronger local contrast, ACM and MDFA still cannot detect all targets. When there are few targets (after-b) or the background is relatively simple (after-c), all algorithms have good performance. For after-d, although the image features have been enhanced, false alarms (ACM, ALC) or missed detections (MDFA) still occur. In addition, both AGPCNet and PFFNet can detect all targets, although PFFNet has a shorter detection time. Obviously, PFFNet is more suitable for drone target detection.

3.4. Ablation Experiments

To investigate the effectiveness of each module in PFFNet, we conducted multiple ablation studies using SIRST Aug and IRSTD 1k. Although ResNet-18 has better performance, it has a longer training and inference time. Therefore, we choose Swin Transformer v1, which performed well and had a shorter detection time as the backbone for ablation.

In Table 2, ablation experiments were conducted on Segmentation Head, FSM, and PFM, respectively. First, for Swin Transformer v1, after adding the Segmentation Head module, the IoU increased by 49.1% (71.3 vs. 47.8) and 49.3% (60.3 vs. 40.4). It can prove the effectiveness of using different downsampling output encoding and multi-scale fusion when segmenting targets. To verify the effectiveness of other modules, we use the Backbone and Segmentation Head modules as baselines to fine-tune the model. After adding the FSM module, the IoU increased by 3% (73.4 vs. 71.3) and 5% (63.4 vs. 60.3). Therefore, the detection accuracy could be improved by enhancing the information exchange between high-level and low-level features. After adding the PFM module, the IoU increased by 0.52% (71.6 vs. 71.3) and 1.8% (61.3 vs. 60.3). Although the texture information of the top-level target is more abundant, there is a lack of effective global contextual priors. Using PFM modules to enhance global feature representation can achieve better performance.

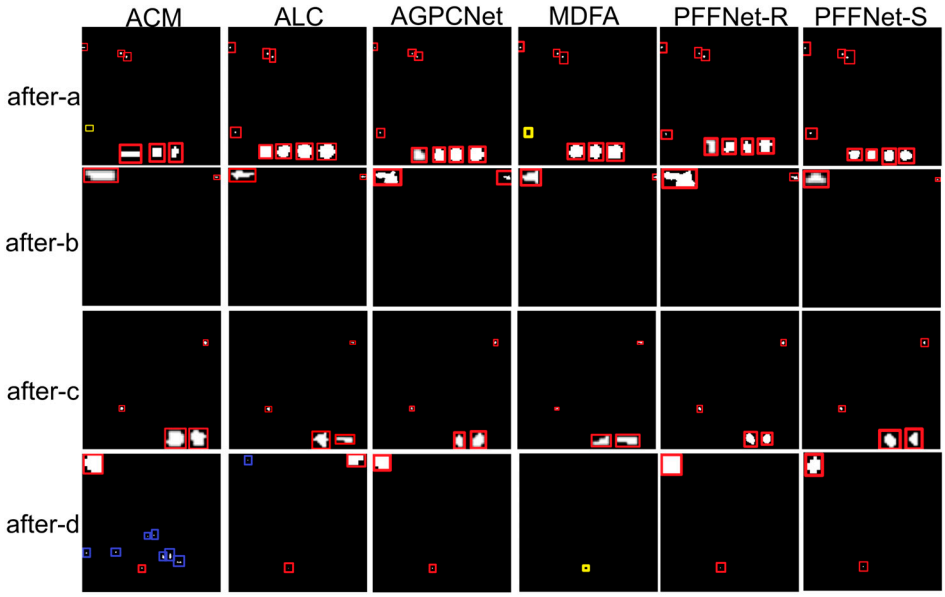


Figure 8. Visual results of different methods. The close-up attempts to display the zoomed-in targets or detection results. Boxes in red, yellow and blue represent correctly detected targets, miss detected targets and false detected targets, respectively. After-(a,b,c,d) indicates the image (a,b,c,d) after enhancement.

Table 2. Ablation of the Segmentation Head, FSM and PFM.

Segmentation Head	FSM	PFM	SIRST Aug(%)		IRSTD 1k(%)	
			IoU	F1-Score	IoU	F1-Score
			47.5	55	40.4	48.7
√			71.3	82.4	60.3	75.3
√	√		**73.4**	**84.7**	**63.4**	**77.6**
√		√	71.6	83.1	61.3	76

For FSM, as shown in Table 3. CSM is used to directly enhance the response weight of targets of the upsampling layer. Although good results can be achieved (71.3 vs. 72.5, 62 vs. 60.3), effective use of lower layer feature maps can achieve better results (72.8 vs. 71.3, 62.6 vs. 60.3). This is because the lower sampling output contains more information of target location and drone contour. Therefore, using different strategies for different sampling output results is beneficial for drone target detection.

Table 3. Ablation of the CSM and LSM in FSM.

CSM	LSM	SIRST Aug (%)				IRSTD 1k (%)			
		IoU	F1-Score	Precison	Recall	IoU	F1-Score	Precison	Recall
		71.3	82.4	77.1	88.6	60.3	75.3	69.8	81.7
√		72.5	83.6	78.6	90.3	62	76.5	72.4	81.2
	√	72.8	84.3	78	91.7	62.6	77.3	75.5	79.1

Dimensionality reduction can reduce redundant information and greatly accelerate the training speed in the network. However, this may result in the loss of useful information. In PFFNet, two dimensionality reductions were performed in FSM and PFM separately, and the reduction ratios were denoted as (r_f, r_p). We choose different dimensionality reduction ratios to explore the best way to segment small infrared targets. According to [38], we set $r_f = 8$. As shown in Table 4, the best result was IoU (73.7%,64.2%) when $r_f = 8$ and $r_p = 4$.

Table 4. Ablation on Dimensionality Reduction Ratio.

Reduction Ratios	SIRST Aug (%)				IRSTD 1k (%)			
	IoU	F1-Score	Precision	Recall	IoU	F1-Score	Precision	Recall
(8,1)	71.8	83.6	77.8	90.3	63.8	77.9	76.6	79.2
(8,2)	73.1	84.4	80.1	89.3	64	78.1	75.7	80.6
(8,4)	73.7	84.8	81	89	64.2	78.2	78.4	77.9
(8,8)	70.6	82.8	74.8	92.7	62.8	77.1	74.2	80.3

With the network deepening and the number of channels increasing, more complex features are learned. In order to enhance the network's expressive ability, it is necessary to cover as many key features as possible. However, it does not apply to all kinds of target detection, such as small drone targets, which have small size and low signal-to-noise ratio. After multiple downsampling, there may be repeated features or even feature loss. Setting a smaller ratio of dimensionality reduction becomes meaningless, such as r_p = 1 or 2. Alternatively, choosing a larger ratio of dimensionality reduction may result in the model being unable to learn more complex features, such as r_p = 8, which is clearly not what we expected. Therefore, an appropriate proportion with different kinds of targets should be chosen to enhance the expressive ability of the model.

4. Discussion

From the above experiments, MDFA divides the generator into miss detection and false alarm and uses GAN to detect infrared targets. However, the results of the downsampling cannot be used effectively. Additionally, MDFA can easily lead to false alarms and miss-detection when the local contrast is low. ACM was the first to use segmentation methods to detect infrared targets, providing a theoretical basis and experimental data for researchers. However, it has many limitations, such as the global information could not be used effectively. ALC combines data-driven and model-driven methods to achieve high precision. However, it ignored the effect of the target's surrounding environment because of only enlarged the scale and enhanced the local contrast. AGPCNet used attention modules and self-attention-guided algorithms to learn more contextual information. However, it can easily lead to miss detection when the target is in a complex environment. Overall, the above methods do not balance the detection accuracy and time cost. Table 5 shows the comparison of all methods.

Table 5. Advantages and disadvantages of all methods (H-high, L-low).

Methods	Advantage	Disadvantage
ACM	L-time	H-FA&MD, L-accuracy
ALC	H-accuracy	H-time, L-recall
MDFA	H-robustness	H-FA&MD, H-time
AGPCNet	H-recall	H-time, L-accuracy
Ours	H-robustness and accuracy	L-FA

In our method, ablation was used to verify the role of each module. For drone targets, the segmentation head was designed to achieve progressive multi-scale fusion. In addition, the problem of feature disappearance was solved. Then, considering the influence of global context prior on the top layer output of the network, the PFM was used to enhance drone feature representation in networks. Finally, FSM was designed to promote information exchange between layers to improve detection accuracy.

5. Conclusions

This paper proposes a fast detection method for infrared small targets: Progressive Feature Fusion Network (PFFNet). The FSM was used to promote the use of shallow

features and enhance upsampling output amplitude of the target. Then, the high-low level encoding outputs are fused to improve the accuracy of the model. The PFM was used to process the output coding diagram of the highest downsampling layer in the backbone and provide more effective global contextual priors for pixel-level scene parsing, guiding algorithms to learn more target features. Then, from the perspective of multi-scale fusion, we design a lightweight segmentation head to adapt to drone targets based on FPN. Enhance the ability to detect small targets by progressively fusing low-level and high-level semantics. Comparative and Ablation experiments demonstrate that PFFNet has the ability to accurately detect drones in complex scenes. Additionally, it can balance the detection accuracy and detection time. This study could promote drone target detection and even expand drone remote sensing segmentation.

However, there are still some problems with the algorithm that need further research, such as dealing with network overfitting, utilizing more efficient contextual information, etc. In future work, RGB images and fusion images will continue to be explored for their application in infrared drone 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. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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. [CrossRef]
2. Li, B.; Song, C.; Bai, S.; Huang, J.; Ma, R.; Wan, K.; Neretin, E. Multi-UAV Trajectory Planning during Cooperative Tracking Based on a Fusion Algorithm Integrating MPC and Standoff. *Drones* **2023**, *7*, 196. [CrossRef]
3. Zhao, M.; Cheng, L.; Yang, X.; Feng, P.; Liu, L.; Wu, N. TBC-Net: A Real-Time Detector for Infrared Small Target Detection Using Semantic Constraint. *arXiv* **2019**, arXiv:2001.05852.
4. Hou, Q.; Wang, Z.; Tan, F.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust Infrared Small Target Detection Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
5. Liao, K.-C.; Wu, H.-Y.; Wen, H.-T. Using Drones for Thermal Imaging Photography and Building 3D Images to Analyze the Defects of Solar Modules. *Inventions* **2022**, *7*, 67. [CrossRef]
6. Li, B.; Yang, Z.-P.; Chen, D.-Q.; Liang, S.-Y.; Ma, H. Maneuvering target tracking of UAV based on MN-DDPG and transfer learning. *Def. Technol.* **2021**, *17*, 457–466. [CrossRef]
7. Fernández, A.; Usamentiaga, R.; de Arquer, P.; Fernández, M.; Fernández, D.; Carús, J.L.; Fernández, M. Robust Detection, Classification and Localization of Defects in Large Photovoltaic Plants Based on Unmanned Aerial Vehicles and Infrared Thermography. *Appl. Sci.* **2020**, *10*, 5948. [CrossRef]
8. Chen, F.; Gao, C.; Liu, F.; Zhao, Y.; Zhou, Y.; Meng, D.; Zuo, W. Local Patch Network with Global Attention for Infrared Small Target Detection. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *58*, 3979–3991. [CrossRef]
9. Ying, X.; Wang, Y.; Wang, L.; Sheng, W.; Liu, L.; Lin, Z.; Zhou, S. Local Motion and Contrast Priors Driven Deep Network for Infrared Small Target Superresolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5480–5495. [CrossRef]
10. Chen, Y.; Li, L.; Liu, X.; Su, X. A Multi-Task Framework for Infrared Small Target Detection and Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–9. [CrossRef]
11. 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*, 13. [CrossRef]
12. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense Nested Attention Network for Infrared Small Target Detection. *IEEE Trans. Image Process.* **2022**, *32*, 1745–1758. [CrossRef]

13. Bai, X.; Zhou, F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognit.* **2010**, *43*, 2145–2156. [CrossRef]
14. Chang, B.; Meng, L.; Haber, E.; Ruthotto, L.; Begert, D.; Holtham, E. Reversible Architectures for Arbitrarily Deep Residual Neural Networks. *Proc. Conf. AAAI Artif. Intell.* **2018**, *32*, 2811–2818. [CrossRef]
15. Rivest, J.; Fortin, R. Detection of dim targets in digital infrared imagery by morphological image processing. *Opt. Eng.* **1996**, *35*, 1886–1893. [CrossRef]
16. Chen, C.L.P.; Li, H.; Wei, Y.; Xia, T.; Tang, Y.Y. A Local Contrast Method for Small Infrared Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 574–581. [CrossRef]
17. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [CrossRef]
18. 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 Geosci. Remote Sens. Lett.* **2014**, *11*, 2168–2172. [CrossRef]
19. 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 Geosci. Remote Sens. Lett.* **2019**, *17*, 1822–1826. [CrossRef]
20. Zhang, L.; Peng, Z. Infrared Small Target Detection Based on Partial Sum of the Tensor Nuclear Norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
21. Zhu, H.; Liu, S.; Deng, L.; Li, Y.; Xiao, F. Infrared Small Target Detection via Low-Rank Tensor Completion with Top-Hat Regularization. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 1004–1016. [CrossRef]
22. 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 Phys. Technol.* **2017**, *81*, 182–194. [CrossRef]
23. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. *arXiv* **2019**, arXiv:1809.02983.
24. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *arXiv* **2019**, arXiv:1911.08287. [CrossRef]
25. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. *Eur. Conf. Comput. Vis.* **2016**, *9905*, 21–37. [CrossRef]
26. 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. [CrossRef]
27. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. *arXiv* **2018**, arXiv:1803.01534.
28. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric Contextual Modulation for Infrared Small Target Detection. *arXiv* **2020**, arXiv:2009.14530.
29. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional Local Contrast Networks for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [CrossRef]
30. Zhang, T.; Cao, S.; Pu, T.; Peng, Z. AGPCNet: Attention-Guided Pyramid Context Networks for Infrared Small Target Detection. *arXiv* **2021**, arXiv:2111.03580.
31. Wang, H.; Zhou, L.; Wang, L. Miss Detection vs. False Alarm: Adversarial Learning for Small Object Segmentation in Infrared Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8508–8517. [CrossRef]
32. 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. [CrossRef]
33. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. *arXiv* **2020**, arXiv:1911.11907.
34. 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.
35. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. *arXiv* **2018**, arXiv:1807.11164. [CrossRef]
36. 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.
37. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2018**, arXiv:1608.06993.
38. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
39. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 6230–6239. [CrossRef]
40. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 936–944. [CrossRef]

41. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape Matters for Infrared Small Target Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 867–876. [CrossRef]
42. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
44. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.