remote sensing

MDPI

Article

AFFPN: Attention Fusion Feature Pyramid Network for Small Infrared Target Detection

Zhen Zuo, Xiaozhong Tong *, Junyu Wei, Shaojing Su, Peng Wu, Runze Guo and Bei Sun

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; z.zuo@nudt.edu.cn (Z.Z.); yujy@nudt.edu.cn (J.W.); ssjing@nudt.edu.cn (S.S.); pengwu@nudt.edu.cn (P.W.); guorunze14@nudt.edu.cn (R.G.); sunbei08@nudt.edu.cn (B.S.)
* Correspondence: tongxiaozhong@nudt.edu.cn

Abstract: The detection of small infrared targets lacking texture and shape information in the presence of complex background clutter is a challenge that has attracted considerable research attention in recent years. Typical deep learning-based target detection methods are designed with deeper network structures, which may lose targets in the deeper layers and cannot directly be used for small infrared target detection. Therefore, we designed the attention fusion feature pyramid network (AFFPN) specifically for small infrared target detection. Specifically, it consists of feature extraction and feature fusion modules. In the feature extraction stage, the global contextual prior information of small targets is first considered in the deep layer of the network using the atrous spatial pyramid pooling module. Subsequently, the spatial location and semantic information features of small infrared targets in the shallow and deep layers are adaptively enhanced by the designed attention fusion module to improve the feature representation capability of the network for targets. Finally, high-performance detection is achieved through the multilayer feature fusion mechanism. Moreover, we performed a comprehensive ablation study to evaluate the effectiveness of each component. The results demonstrate that the proposed method performs better than state-of-the-art methods on a publicly available dataset. Furthermore, AFFPN was deployed on an NVIDIA Jetson AGX Xavier development board and achieved real-time target detection, further advancing practical research and applications in the field of unmanned aerial vehicle infrared search and tracking.

Keywords: small infrared target detection; attention fusion; atrous spatial pyramid pooling; feature pyramid network

Citation: Zuo, Z.; Tong, X.; Wei, J.; Su, S.; Wu, P.; Guo, R.; Sun, B. AFFPN: Attention Fusion Feature Pyramid Network for Small Infrared Target Detection. *Remote Sens.* **2022**, *14*, 3412. https://doi.org/10.3390/rs14143412

Academic Editors: Senthilnath Jayavelu, Yongshuo Fu and Mohammad Rostami

Received: 19 June 2022
Accepted: 13 July 2022
Published: 15 July 2022

1. Introduction

Small infrared target detection technology, which has been widely used in several fields, such as military early warning, precision guidance, field rescue, and forest fire prevention, forms an essential part of an infrared search and tracking system [1]. Compared to other imaging methods, infrared imaging offers a longer range, resistance against interference, and independence from lighting conditions. However, the detected targets are usually very small—ranging from one pixel to tens of pixels—with weak texture, shape information, and low signal-to-noise ratios, owing to the long distance between the target and the infrared sensor. These targets tend to become lost in the presence of heavy noise and background clutter (Figure 1). The unique characteristics of small infrared target imaging pose significant technical challenges; therefore, the accurate and effective detection of small infrared targets remains an open problem.

For decades, small infrared target detection has predominantly been based on model-driven conventional methods. These methods, which analyze the physical and imaging properties of the target, make reasonable assumptions based on prior knowledge, and design fixed hyperparameters owing to the lack of publicly available infrared datasets. Model-driven methods primarily include background suppression-based methods [2,3],

local contrast-based methods [4–7], and optimization-based methods [8–10]. Despite the abundance of theoretical and hypothetical constraint terms, these model-driven methods exhibit low detection accuracy and poor robustness in practical detection tasks when the target size, target shape, signal-to-noise ratio, and background clutter significantly change.

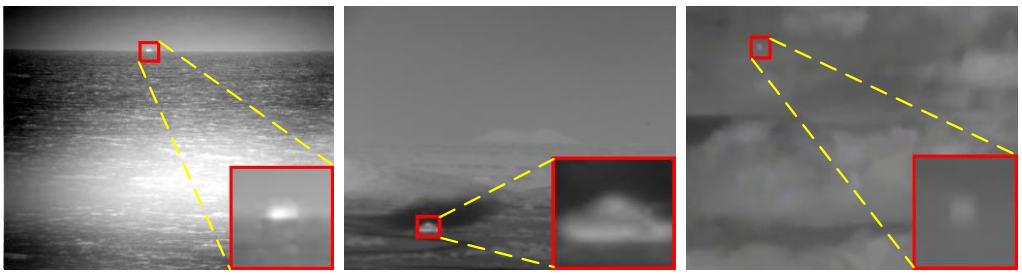


Figure 1. Example images of small infrared targets, indicated by the red bounding box and magnified in the lower-right corner. Left: a ship that is difficult to identify against a complex background of sea clutter; middle: a tank that is difficult to identify in a mountainous scene; right: a dim five-pixel-sized target against a cloudy background.

In recent years, publicly available infrared datasets [11,12] have fueled the development of data-driven methods for small infrared target detection. Unlike traditional model-driven approaches, convolutional neural network (CNN)-based methods train detectors with a large amount of data to enable models to learn the features of the target; consequently, they have significant advantages in terms of accuracy and false alarm. Gao et al. [13] proposed an approach based on a constant false alarm rate to detect dim and small targets. McIntosh et al. [14] proposed an optimized target-to-clutter ratio metric for small infrared target detection networks. Du et al. [15] demonstrated that in small infrared target detection tasks, the focus should be on shallow features containing rich detail and spatial location information. Zhao et al. [16] used a generative adversarial network (GAN) to detect small infrared targets, but this approach requires numerous training samples and is time-consuming. Dai et al. [17] focused on feature fusion in different network layers and designed the asymmetric contextual module (ACM) specifically for small infrared target detection. This enabled them to obtain a more effective feature representation, thus improving the detection of small infrared targets. Chen et al. [18] proposed a local patch network with global attention by considering the global and local characteristics of infrared small targets. Hou et al. [19] constructed a feature extraction framework, combining manual feature methods to build a mapping network between feature maps and small target likelihoods in images. Hou et al. [20] converted a single infrared image frame into a probabilistic likelihood map of the target in terms of the image pixels, introduced feature groups into network downsampling at the perception layer, enhanced the small target feature group weights to improve the representation of small targets, and introduced a skip connection layer with full convolution. Ma et al. [21] proposed a small infrared target detection network with generated labels and feature mapping to deal with the low contrast and low signal-to-noise ratio characteristics of small infrared targets. Wang et al. [22] focused on the correlation between target and background pixels and proposed a coarse-to-fine internal attention-aware network for small infrared target detection. Wang et al. [23] proposed a multi-patch attention network based on an axial attention encoder and a multi-scale patch branching structure in order to highlight the effective characteristics of small targets and suppress background noise. Chen et al. [24] proposed a multi-tasking framework for infrared small target detection and segmentation that reduces the model complexity while significantly improving the inference speed of the algorithm. Zhou et al. [25] proposed the competitive game framework pixelgame from a new perspective by highlighting the target information through maximum information modulation. Although these data-driven methods have employed various techniques to achieve detection performance gains, they are affected by the target being lost in the deep layers and poor edge segmentation details.

Based on the above analysis, for features such as missing color information and the weak texture shape of small infrared targets, the target detection network needs not only to fully utilize the global contextual information of the image but also to add attention mechanisms to focus on the target regions of interest in different network layers. Inspired by [26–28], we designed the attention fusion feature pyramid network (AFFPN) for small infrared target detection. Without increasing the model complexity, the network focuses on the important spatial location and channel information of small targets through attention fusion and the acquisition and exploitation of image global contextual information to enhance the feature representation of small targets, thus improving the detection performance.

The contributions of this study are summarized as follows:

(1) We propose the AFFPN for single-frame small infrared target detection, which achieves a better performance than existing methods on the publicly available SIRST dataset, enabling effective segmentation of small target details, and exhibits higher robustness against complex backgrounds.
(2) We propose an attention fusion module that focuses on the channel and spatial location information of different layers and uses global contextual information to achieve feature fusion. This module helps the network focus on the semantic and detailed information of the infrared mini-target and dynamically perceives the features of the different network layers of small targets.
(3) We deploy the proposed algorithm on an NVIDIA Jetson AGX Xavier development board and achieve real-time detection of 256 $\times$ 256-pixel resolution images.

The remainder of this paper is organized as follows. We begin with a brief review of the related work in Section 2. Subsequently, in Section 3, we provide a detailed description of the AFFPN structure. Section 4 presents the experimental details and analyzes the results obtained. Finally, the conclusions are presented in Section 5.

2. Related Work

2.1. Small Infrared Target Detection

Several studies have focused on small infrared target detection [1]. Traditional model-driven approaches rely on reasonable assumptions, require no data training, and the detection results focus on the location of the target in the image. Filter-based approaches measure the discontinuity between the target and the background and include top-hat filters [29] and max-median filters [3]. Human visual system-based approaches [4,5,30,31] detect small targets based on local contrast differences between the target and the background, extracting valid feature information from complex backgrounds. However, these model-driven approaches are susceptible to factors such as background clutter and noise, which degrade the detection performance.

In recent years, CNNs have been employed for small infrared target detection as a pixel-level segmentation task and have attracted considerable attention. The CNN-based approach has powerful feature representation capabilities and achieves a better detection performance than conventional methods in terms of learning small target features from large amounts of data. Deng et al. [32] proposed a multiscale CNN for spatial infrared point target recognition, which combines shallow and deep features for feature learning and classification. Shi et al. [33] transformed the detection problem into a denoising problem and achieved end-to-end detection of small infrared targets. Additionally, [13,14,34] designed CNNs with different structures to extract the basic features of small infrared targets to improve the detection accuracy. Zhao et al. [35] proposed a lightweight small infrared target detection network called TBC-Net, which used U-Net [36] as the target extraction module and designed a semantic constraint module. Zhao et al. [16] constructed a GAN model to automatically learn the unique distribution features of small infrared targets and directly predict the intensity of the targets. Attention mechanisms help networks focus on regions of interest and enhance the contextual information of target features [15]. Dai et al. [17] proposed an asymmetric attention module specifically for small infrared target detection. They designed top-down global context feedback and bottom-up modulation paths to

exchange deep semantic information and fine shallow detail information to better detect small targets. Li et al. [27] proposed the densely nested attention network (DNANet) and designed a dense nested interaction module (DNIM) for multiple interactions between deep semantic features and shallow detail features to maintain deep small infrared targets. Zhang et al. [37] focused on the contextual relationships and feature utilization in network delivery for small infrared target detection and proposed the attention guided pyramidal context network (AGPCNet), which improves the detection accuracy through shallow and deep feature fusion.

2.2. Attention and Feature Fusion

Attention mechanisms: Attention mechanisms have received considerable attention owing to their excellent performance and have been widely used in target detection, semantic segmentation, and natural language processing. Self-attention [38] is a popular attention mechanism that extracts attention based on the attention of the feature map; non-local attention convolutional units [39] focus only on regions that are the size of the neighborhood kernel; SENet [26] automatically obtains the importance of each channel and performs weighted selection by the interdependencies between the modeled features displayed; CBAM [40] considers the attention weights in both the spatial and channel dimensions to better focus on regions of interest; DANet [41] uses self-attention to fuse channel attention (CA) and spatial attention (SA); and [42–45] focus on the region of interest of the image from different perspectives to obtain excellent enhancement effects.

Feature fusion: The emergence of U-Net [36] and the feature pyramid network (FPN) [46] has boosted the development of semantic segmentation networks, enabling better feature fusion between shallow and deep networks using summation or concatenation between different levels. Zhang et al. [47] designed the semantic feature fusion module to solve the information imbalance caused by information dilution in an FPN. Gong et al. [48] designed the fusion factor to control the information passed from the deep to shallow layers to adapt the FPN to tiny object detection. Li et al. [27] proposed the DNIM to achieve the fusion of deep semantic features and shallow detailed features while constructing an attention mechanism between different layers to further enhance the fusion performance. Tong et al. [49] proposed an enhanced asymmetric attention feature fusion module that preserves and highlights the fine details of small infrared targets using shuffle units and cross-layer feature fusion. Zhang et al. [37] focused on attention-guiding mechanisms and fusion strategies between contexts for small infrared target detection. Attention mechanisms and feature fusion strategies in neural networks have evolved rapidly; however, these studies have focused on generic target detection. Studies on semantic segmentation tasks and architectures specifically for small infrared target detection tasks are limited. The low signal-to-noise ratio of small infrared target images, small number of image pixels constituting the target, lack of texture information, absence of contour and shape features, and complex background represent the key challenges involved in small infrared target detection. Therefore, developing and designing effective attention modules and feature fusion mechanisms for small infrared target detection is crucial.

3. Proposed Method

In this section, we introduce the details of the proposed AFFPN. The network architecture of the proposed method is shown in Figure 2. The network consists of two parts: the feature extraction and feature fusion modules. The following subsections describe the overall structure and main component modules of the proposed AFFPN.

3.1. Network Architecture

Figure 2 shows that, given the SIRST image as input, the AFFPN performs feature extraction and feature fusion in turn, and then the prediction module classifies each pixel and outputs the segmentation result of the small infrared target. Section 3.2 presents details on the feature extraction module, which consists of two parts: the atrous spatial

pyramid pooling module (ASPPM), which acquires global contextual information, and the attention fusion (AF) module, which fuses different feature layers. The input image is pre-processed and downsampled to extract rich small target features, and the global contextual information of the small targets is aggregated and utilized by the ASPPM. The attention fusion module enhances the network's feature representation of small infrared targets by skipping connections and passing high-resolution information throughout the network, ensuring effective fusion of spatially detailed shallow features and semantically rich deep features. We use the attention fusion module to adaptively enhance deep and shallow features to obtain more accurate channel and spatial location information of the target because a gap exists between the different layers in terms of semantic information.

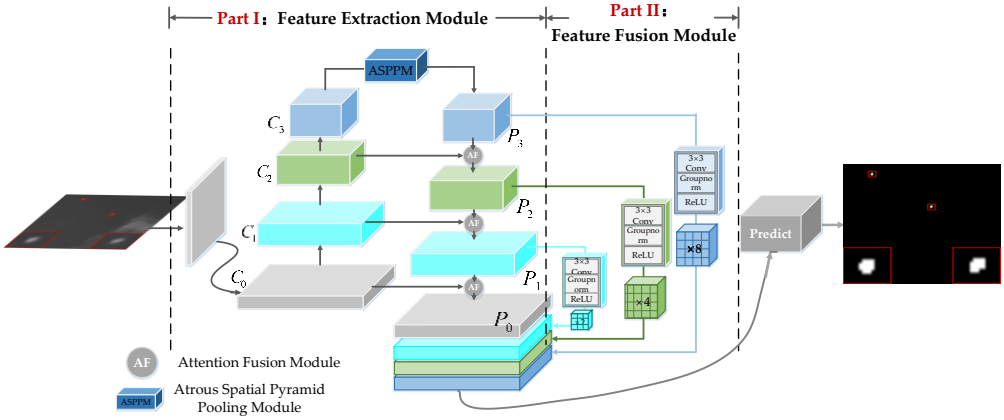


Figure 2. Proposed AFFPN. Feature extraction module. First, the input image is sent to the residual module for downsampling feature extraction and then the atrous spatial pyramid pooling and attention fusion modules for selective enhancement of features at different levels. Feature fusion module. The features of different levels are upsampled and concatenated to fuse the multilayer features, and the segmentation result of the target is obtained by the prediction module.

Notably, we use ResNet-20 [50] as the backbone architecture in the C_1, C_2, and C_3 phases of feature extraction, with the details of the network backbone shown in Table 1 (only the first convolutional layers of C_2 and C_3 were subsampled), enhancing the learning capability of the CNN and further improving the network's ability to make full use of features at different levels during the downsampling and upsampling stages. To ensure that the spatial location information at the shallow level is passed to the deeper network layers to reduce the loss of fine detail information at the shallow level for small infrared targets at the deeper levels, the feature fusion module is described in detail in Section 3.3. Different feature layers are upsampled to the same size and then the features of the different network layers are fused by a concatenation operation to generate a robust feature map capable of improving the feature representation of small targets. The final binary map output by the prediction module is the small infrared target detection result.

Table 1. AFFPN backbones.

Stage	Output	Backbone
C_0	480 × 480	$3 \times 3\text{conv}, 64$
C_1	480 × 480	$\begin{bmatrix} 3 \times 3\text{conv}, 64 \\ 3 \times 3\text{conv}, 64 \end{bmatrix} \times 3$
C_2	240 × 240	$\begin{bmatrix} 3 \times 3\text{conv}, 128 \\ 3 \times 3\text{conv}, 128 \end{bmatrix} \times 3$
C_3	120 × 120	$\begin{bmatrix} 3 \times 3\text{conv}, 256 \\ 3 \times 3\text{conv}, 256 \end{bmatrix} \times 3$

3.2. Feature Extraction Module

3.2.1. ASPPM

Zhao et al. [51] demonstrated that pyramidal pooling modules can construct effective global contextual priors to reduce the loss of contextual information. We constructed the hierarchical contextual ASPPM in the deepest feature layer (C_3) of the FPN infrastructure to fully utilize the global information, as shown in Figure 3. We believe that a single 3×3 or 1×1 convolution mixing all multi-scale contextual information is insufficient. Inspired by [51], we first increase the perceptual field of the deep network by atrous convolution with three different dilation rates, using three different-sized pooling modules divided into different sub-regions to form pooled representations for different locations. The pooled feature maps are then upsampled and fused with the feature maps of different sizes of atrous convolution to obtain a rich perceptual field in a stitching fashion to form the final feature representation.

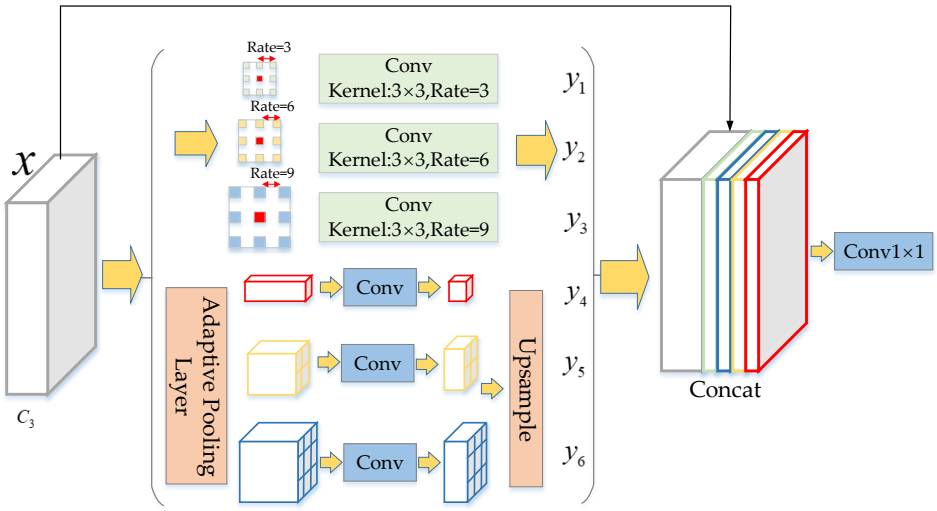


Figure 3. Architecture of the atrous spatial pyramid pooling module (ASPPM).

A hierarchical global prior that includes different scales and variations between sub-regions can reduce the information loss between sub-regions [52]. We propose the novel ASPPM to better capture the global contextual prior information of small infrared target images. For pyramids with a layer size of N, the perceptual field of the network is first increased by convolution of voids with different dilation rates. Then, the dimensionality of the contextual representation is reduced to 1/N of the original layer using a 1×1 convolutional layer, and the feature map is upsampled to the same size as the input features (C_3) by means of bilinear interpolation. Finally, the features of different levels are concatenated to form the final global contextual scene prior. Notably, the size of the atrous convolution dilation rate and the size of the global pooling is changeable. We set the size of the ASPPMs dilation rate to 3, 6, and 9 and the bin sizes of the ASPPMs to different scales of 1×1, 2×2, and 4×4 based on the pixel distribution properties of a small infrared target. The ASPPM of the global contextual prior is calculated as:

$$y_j = \begin{cases} \delta(\mathcal{B}(Conv_{3\times3}(x, dilation = i))) & i = 3,6,9 \qquad j = 1,2,3 \\ U(Conv_{1\times1}(\mathcal{B}(\delta(AdapPool(x,i))))), & i = 1,2,4 \qquad j = 4,5,6 \\ \mathbf{L}(\mathbf{X}) = Conv_{1\times1}(Concat[y_j(x), x]) \end{cases} \tag{1}$$

where $Conv_{1\times1}$ is point-by-point convolution [53] with a kernel size of 1×1 and $Conv_{3\times3}$ is the different dilation rate atrous convolution with a kernel size of 3×3. U, $\mathcal{B}$, and δ are upsampling, batch normalization, and the ReLU activation function, respectively. *Concat* is the operation of cascading features at different scales, and *AdapPool* is the global pooling operation with different feature sizes.

3.2.2. Attention Fusion Module

Deep feature networks extract the semantic information of small targets; however, they may risk losing the spatial details of the targets. Shallow feature networks retain the spatial location information; however, they lack a deep semantic understanding of the targets. The original FPN [46] consists of a bottom-up feedforward network, skip connection, and top-down network. The feedforward network is used to expand the field of perception and extract high-level semantic information, and the inverse network restores high-level features to the same size as the input image. A simple approach for achieving powerful contextual information modeling capability is to continue increasing the depth of the network. However, the areas of small infrared targets consist of only a few pixels and as the network deepens, the target may be lost in the deeper network. Additionally, feature representation is difficult when the network does not make full use of the information from the different feature layers.

Therefore, we propose the attention fusion module (Figure 4) to ensure the extraction of rich semantic information of small targets while maintaining the deep feature representation of the small targets. Moreover, this module is used to solve the problems of target detail loss, information redundancy between different layers, and inadequate feature fusion. Here, X and Y refer to the shallow fine detail information and deep semantic information, respectively. The deep semantic feature map is upsampled to the same size as that of the corresponding shallow feature map, and then the feature map output Z is obtained through the attention fusion module. The attention fusion module consists of SA, which focuses on the shallow spatial location information of the target, and CA, which focuses on the deep semantic information of the target. The input shallow detail features X and deep semantic features Y are processed by the SA and CA, respectively, and then the processed features are fused.

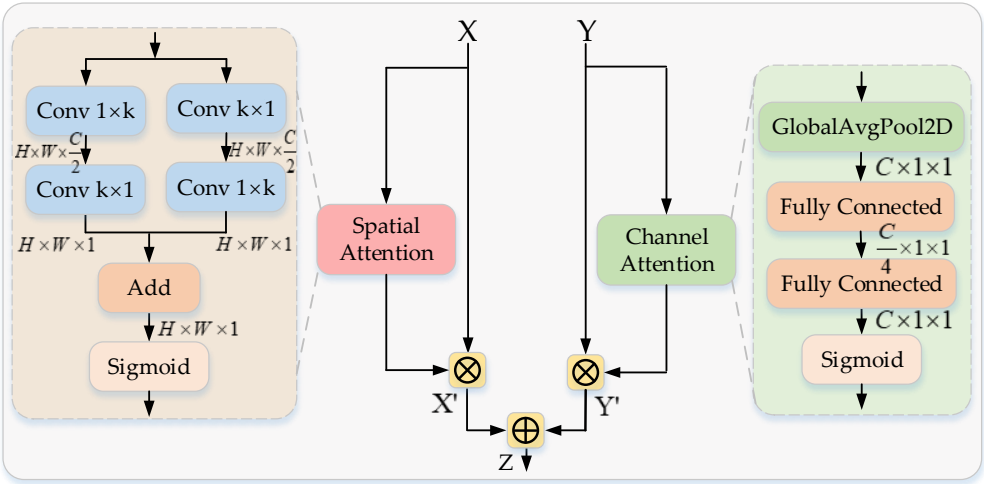


Figure 4. Architecture of the attention fusion module.

SA: The shallow feature map contains a large amount of detailed information [28], and, hence, we use SA to learn the details of small targets in order to direct the network's attention to more detailed spatial location information. This helps to generate effective features for small target detection. We denote the shallow features as $X \in \mathbb{R}^{C\times H\times W}$, and the set of locations is denoted by $\mathbb{R}= \{(x,y)|x = 1,2,\ldots H; y = 1,2,\ldots W\}$, where (x,y) represents the location coordinates of the features. Similar to [28], the input shallow feature maps are convolved using convolution kernel sizes of $1 \times$ k and k $\times$ 1 to, respectively, capture the spatial location and detail information of the shallow features of small infrared targets. Finally, the spatial feature maps are normalized using the sigmoid function. SA

attention weights $S(\mathbf{X}) \in \mathbb{R}^{C\times H\times W}$ that focus on the shallow spatial location information are calculated as:

$$\begin{aligned} C_1 &= Conv_2(Conv_1(X, W_1^1), W_1^2) \\ C_2 &= Conv_1(Conv_2(X, W_2^1), W_2^2) \\ SA = S(X, W) &= \sigma(C_1 + C_2) \end{aligned} \tag{2}$$

where $Conv_1$ and $Conv_2$ are convolutional operations with kernel sizes of $1 \times k$ and $k \times 1$, respectively, and W and the sigmoid function are the SA parameters. Finally, the output of the shallow features is obtained by weighting the SA as:

$$X\prime = \mathrm{SA}(X) \otimes X \tag{3}$$

where $\otimes$ and $\mathrm{SA}(\cdot)$ denote element-wise multiplication and SA, respectively.

CA: Deep features contain highly abstract semantic information [50]; however, they lack detailed information about the target. We use global CA [26] to weight deep features using multiple receptive fields to capture highly discriminative channel features. Channel-level statistics are first generated by Equation (4) to obtain a feature map of size $1 \times 1 \times C$ with global sensory fields, aggregating global contextual information. The dependencies between channels are then captured through two fully connected layers, as shown in Figure 4. CA attention weights $\mathrm{CA}(Y) \in \mathbb{R}^{C\times H\times W}$, which focus on deep channel information, are calculated as:

$$y = \frac{1}{H \times W} \sum_{i=1,j=1}^{H,W} Y[:, i, j] \tag{4}$$

$$\mathrm{CA} = \sigma(\mathcal{B}(W_2\delta(\mathcal{B}(W_1 y)))) \tag{5}$$

where H, W, σ, $\mathcal{B}$, and δ denote the height, width, sigmoid function, batch normalization, and ReLU function, respectively. Moreover, W_1 and W_2 represent two fully connected operations, where $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$, and r denotes the channel reduction ratio. Finally, the output of the deep features is obtained according to the CA weights:

$$Y\prime = \mathrm{CA}(Y) \otimes Y \tag{6}$$

where $\otimes$ and $\mathrm{CA}(\cdot)$ denote element-wise multiplication and CA, respectively. The final output after SA and CA for shallow and deep features, respectively, is the attention feature fusion map:

$$Z = \mathrm{SA}(X) \otimes X + \mathrm{CA}(Y) \otimes Y \tag{7}$$

3.3. Feature Fusion Module

A single feature map path cannot adequately represent the small infrared target features after the feature extraction module. We designed a pyramid feature fusion module specifically for aggregating deep and shallow features to achieve multilayer feature fusion of small infrared targets, as shown in Figure 2. We first upsample the feature maps of different layers to the same size $P_{up}^{i,J} \in \mathbb{R}^{C_i\times H_0\times W_0}$ $i \in \{0, 1, \ldots, I\}$ and concatenate the deep features containing semantic information and shallow features with fine detail information to generate feature maps containing rich information. This prevents feature loss and underutilization of shallow spatial location information in the deep layer network for small targets and ensures the robustness of small target features. The feature map generation P_i' for multi-feature fusion is obtained using the following equation:

$$P_i' = U_{2^i}(\delta(GN(Conv_{3\times3}(P_i)))) \qquad \{i = 1, 2, 3\} \tag{8}$$

where $U(\cdot)$ denotes upsampling using bilinear interpolation and 2^i is a multiple of upsampling. $Conv_{3\times3}$, GN, and δ denote convolutional operations with kernel sizes of 3×3, group normalization, and the ReLU function, respectively.

4. Experimental Evaluation

In this section, we present qualitative and quantitative evaluations of the proposed method conducted using publicly available infrared datasets. First, the evaluation metrics are described in Section 4.1, and details of the experimental implementation are described in Section 4.2. Subsequently, we present the results of a detailed ablation study in Section 4.3. Finally, in Section 4.4, the AFFPN is compared visually and quantitatively with state-of-the-art methods to demonstrate its superiority.

4.1. Evaluation Metrics

We consider small infrared target detection as a pixel-level semantic segmentation task. Therefore, we use classical semantic segmentation evaluation metrics to compare the performance of different algorithms. The main algorithm evaluation metrics include mean intersection over union (mIoU), normalized IoU (nIoU), F-measure, the precision–recall (PR) curve, and receiver operating characteristic (ROC). These evaluation metrics assess the ability of the algorithm to accurately locate and describe the shape of small infrared targets to ensure that the network detects the target and to ensure there are as few false positives as possible.

(1) Mean intersection over union (mIoU): mIoU is the classical pixel-level semantic segmentation evaluation metric used to characterize the contour description capability of an algorithm. It is defined as the ratio of the intersection and concatenation area between predictions and labels, as follows:

$$\text{mIoU} = \frac{\text{\# Area of Overlap}}{\text{\# Area of Union}} \tag{9}$$

(2) Normalized IoU (*nIoU*): *nIoU* is an evaluation metric designed by [11] for small infrared target detection to better measure the segmentation performance of small targets and prevent the impact of the segmentation results of large targets on the overall evaluation metric. It is defined as follows, where *TP, T,* and *P* denote true positive, true, and positive, respectively:

$$nIoU = \frac{1}{N}\sum_{i}^{N}\frac{TP[i]}{T[i] + P[i] - TP[i]} \tag{10}$$

(3) F-measure: The F-measure is used to measure the relationship between precision and recall. Precision, recall, and the F-measure are defined as follows, where $\beta^2 = 0.3$, and *FP* and *FN* denote the numbers of false positives and false negatives, respectively:

$$\begin{gathered} Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN} \\ F_{measure} = \frac{(\beta^2+1)Precision \times Recall}{\beta^2 Precision + Recall} \end{gathered} \tag{11}$$

(4) PR curve: The PR curve is used to characterize the dynamic change between precision and recall; the closer the curve is to the upper right, the better the performance. Average precision (*AP*) is used to accurately evaluate the PR curve, as defined as follows, where *P* is precision and *R* is recall:

$$AP = \int P(R)dR \tag{12}$$

(5) ROC: The dynamic relationship between true positive rate (*TPR*) and false positive rate (*FPR*) is described by the ROC. The *TPR* and *FPR* are defined as follows, where *FN* denotes the number of false negatives:

$$TPR = \frac{TP}{TP+FN} \quad FPR = \frac{FP}{FP+TN} \tag{13}$$

The area under the curve (AUC) is used to quantitatively assess the ROC.

4.2. Implementation Details

Dataset: We evaluated the proposed AFFPN on the SIRST [11] dataset, which contains 427 images and 480 typical small infrared target instances. We divided the training, validation, and test sets using a 5:2:3 ratio. Several small infrared targets in the SIRST dataset were very dim and hidden in background clutter, and only 35% of the targets contained the brightest pixels in the image (Figure 1). Therefore, thresholding the original image considering only the salient features of the target or background suppression-based methods does not achieve good detection results.

Implementation details: We conducted experiments for all data-driven methods using implementations based on the PyTorch framework. The input images were randomly cropped to the same size of 480 × 480. All images were normalized to accelerate network convergence. The Adagrad optimizer [54] is widely used for network training, and it is one of the most popular optimizer methods. We refer to Dai et al. [17] for training AFFPN using the Adagrad optimizer and the corresponding hyperparameters, with a batch size of 8 and the initial learning rate and weight decay set to 0.05 and 1×10^{-4}, respectively. The proposed network model was trained using the softIoU loss function [55] for 300 epochs. All model-driven methods were re-evaluated on the SIRST test set using MATLAB 2019. For the data-driven methods (FPN [46], U-Net [36], TBC-Net [35], and ACM-FPN [17]), we re-trained and tested them on the SIRST dataset according to the authors' publicly available code and original parameter settings to obtain comparative experimental results. The TBC-Net [35] experimental results were referenced from Dai et al. [11]. A computer with an Intel I9-10900X @3.7 GHz CPU and a single TITAN RTX GPU was used for training and testing.

4.3. Ablation Study

To demonstrate the effectiveness of the structural design and network modules of the AFFPN, different variants were constructed for detailed ablation experiments.

(1) Ablation study for the attention fusion module: The attention fusion module adaptively enhances shallow spatial location features and deep semantic features, filtering redundant features while focusing on the valuable information of the target in different layers to achieve better feature fusion. We compared AFFPN with four variants to demonstrate the effectiveness of the designed attention fusion module.

- AFFPN-cross-layer feature fusion: We considered cross-layer feature modulation between different feature layers, changing the feature layers that CA and SA focus on. Specifically, the features of the shallow layer are dynamically weighted and modulated by SA and the features of the deep layer, and the features of the deep layer are weighted and modulated by CA and the features of the shallow layer. Finally, their features are summed to fuse them, as shown in Figure 5a.
- AFFPN w/o AF (element-wise summation): This variant of AFFPN removes the CA and SA modules and uses the common element-wise summation approach instead of the AF module to achieve feature fusion in different layers. The aim is to explore the effectiveness of the AF module, as shown in Figure 5b.
- AFFPN w/o SA: We considered only CA in this AFFPN variant, and removed SA to investigate its contributions, as shown in Figure 5c.
- AFFPN w/o CA: We considered only SA in this variant, removing CA to evaluate its advantages, as shown in Figure 5d.

The results of the ablation experiments for different variants are listed in Table 2, where larger values of mIoU, nIoU, and F-measure are associated with a better performance and the opposite is true for the number of parameters. The best results in each column are highlighted in boldface red font and the second-best results are highlighted in boldface blue font. We set up an ablation module for cross-layer feature fusion to explore the effect

of information interaction between the shallow detail features and deep semantic features. The results in the table show that the mIoU, nIoU, and F-measure of AFFPN–cross-layer feature fusion decreased by 2.25%, 1.7%, and 1.13%, respectively. The experimental results show that maximizing the advantages of different feature layers is key to improving the feature representation capability of the network for small infrared target detection.

Table 2. Different attention-fusion modules and their mIoU, nIoU, and F-measure results on the SIRST dataset.

Model	Params (M)	mIoU ($\times 10^{-2}$)	nIoU ($\times 10^{-2}$)	F-Measure ($\times 10^{-2}$)
AFFPN–cross-layer feature fusion	7.40	75.89	74.21	82.50
AFFPN w/o AF (element-wise summation)	**7.17**	75.80	**74.63**	82.48
AFFPN w/o SA	**7.18**	**76.26**	74.43	82.01
AFFPN w/o CA	7.39	75.58	74.15	**82.93**
AFFPN (Ours)	7.40	**78.14**	**75.91**	**83.63**

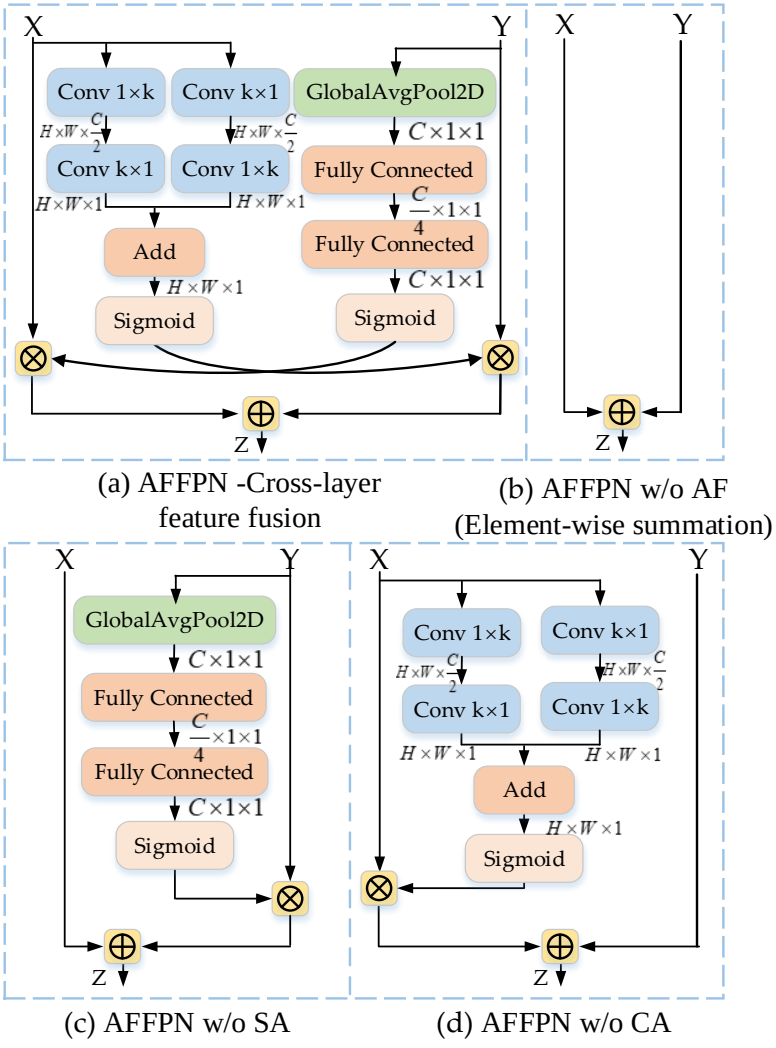


Figure 5. Architecture for attention to fusion module ablation studies.

The table shows that, if the AF module is removed, the mIoU, nIoU, and F-measure of the AFFPN w/o AF on the test set decrease by 2.34%, 1.28%, and 1.15%, respectively. This

demonstrates the importance of SA and CA fusion, where the AF module pays sufficient attention to the spatial location and semantic feature information of different layers and fuses the information of different sub-features to enhance the feature representation of CNN for small infrared targets, thus enhancing the performance of small infrared target segmentation. To further illustrate the effectiveness of our proposed attention fusion module, we visualized the heat map before and after attention fusion. As shown in Figure 6, with the help of the attention fusion module, the AFFPN deep feature maps are accurately shape segmented and have a strong response to information cues.

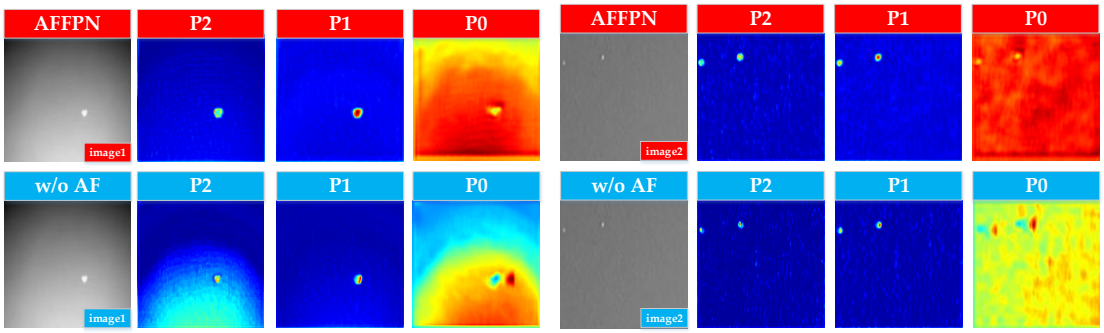


Figure 6. Visualization map of AFFPN and AFFPN w/o attention fusion. The output of AFFPN is circled by a solid red frame. The feature maps from the deep layer of AFFPN have high values representative of informative cues.

Furthermore, we conducted separate ablation experiments with AFFPN variants that lack SA and CA to explore the contribution of the two different types of attention. As Table 2 shows, there is a significant decrease in the mIoU, nIoU, and F-measure scores of both AFFPN w/o SA and AFFPN w/o CA. This is because SA helps the network focus on shallow fine detail features and location information, and CA better utilizes deep semantic information to enhance the network representational capability, thus producing better small target detection results.

Additionally, to demonstrate the novelty and effectiveness of our attentional feature fusion module designed for small infrared targets, we compared it with typical attentional fusion approaches incorporating the attentional mechanisms SENet [26], CBAM [40], and Shuffle Attention [45].

The experimental results in Table 3 further demonstrate the effectiveness of our designed attention fusion approach. The designed attention fusion mechanism adds little to no network complexity and exhibits the best performance on all evaluation metrics. Instead of simply superimposing attention, we focus on the spatial features details of the shallow layers and the semantic features of the deeper layers, concentrating on the representative features of the different layers of the target, thereby significantly improving the detection performance of small infrared targets. The experimental results demonstrate the novelty and effectiveness of our idea of attentional fusion focusing on different layers of target features.

Table 3. Various attention modules and their mIoU, nIoU, and F-measure results on the SIRST dataset.

Model	Params (M)	mIoU ($\times 10^{-2}$)	nIoU ($\times 10^{-2}$)	F-Measure ($\times 10^{-2}$)
AFFPN with SE	**7.40**	**76.71**	**75.06**	82.60
AFFPN with CBAM	**7.28**	75.97	74.02	**82.90**
AFFPN with Shuffle Attention	7.46	74.47	73.11	82.56
AFFPN (Ours)	**7.40**	**78.14**	**75.91**	**83.63**

(2) Ablation study for ASPPM and multiscale feature fusion: The ASPPM is used to enhance the global a priori information of a target and reduce contextual information loss. Multiscale feature fusion concatenates deep features containing semantic information and shallow features containing spatial location detail information to generate globally robust feature maps in order to improve the detection performance of small targets. We compared AFFPN with two variants to demonstrate the effectiveness of ASPPM and multiscale feature fusion.

- AFFPN w/o ASPPM: We removed the ASPPM from this variant to assess its contribution.
- AFFPN w/o multilayer concatenation: We removed the multilayer feature fusion module in this variant and used the last layer of the feature extraction module to predict the targets to explore the effectiveness of multiscale feature fusion.

Table 4 shows the experimental results obtained for the AFFPN and its variants. The AFFPN without the ASPPM shows degradation on all metrics. We improved the performance of small infrared target detection with respect to all metrics, with almost no increase in the number of network parameters because we designed the ASPPM to reduce contextual information loss and better capture the global contextual a priori information of the target.

Table 4. ASPPM and multiscale feature fusion ablation results for mIoU, nIoU, F-measure, AP, and AUC.

Model	Params (M)	mIoU ($\times 10^{-2}$)	nIoU ($\times 10^{-2}$)	F-Measure ($\times 10^{-2}$)	AP ($\times 10^{-2}$)	AUC ($\times 10^{-2}$)
AFFPN w/o ASPPM	**7.63**	**76.32**	**74.59**	83.29	**79.17**	**94.44**
AFFPN w/o multilayer concatenation	7.68	74.25	74.55	**83.42**	78.53	93.67
AFFPN (ours)	**7.40**	**78.14**	**75.91**	**83.53**	**80.61**	**94.52**

Compared to the results of the AFFPN, the mIoU, nIoU, F-measure, AP, and AUC values of the AFFPN without multilayer concatenation decreased by 3.89%, 1.36%, 0.11%, 2.08%, and 0.85%, respectively. This is due to the limited features available in a single feature layer, which fails to fully utilize the rich semantic information in the deep layer and the fine detail information in the shallow layer. This further demonstrates the rationality and superiority of using multiscale features for fusion in the proposed feature fusion part of AFFPN.

4.4. Comparison with State-of-the-Art Methods

To further demonstrate the superiority of AFFPN, we performed qualitative and quantitative comparisons with state-of-the-art methods. First, eight traditional model-driven methods were selected for comparison with AFFPN. They included the filter-based methods top-hat [29] and max-median [3], the human visual system-based methods RLCM [7] and MPCM [56], the gradient property-based LIGP [57], the multiscale image entropy-based MGDWE [58], and the optimization-based methods NRAM [59] and PSTNN [60]. The parameter settings for these model-driven methods are shown in Table 5. Additionally, FPN [46], U-Net [36], TBC-Net [35], ACM-FPN [17], and ACM-U-Net [17] were selected as data-driven comparison methods.

Table 5. Parameter settings of the model-driven methods.

Methods	Parameter Settings
Top-hat	Structure size = 3×3
Max-median	Patch size = 3×3
RLCM	Size: 8×8, Slide step: 4, threshold factor: k = 1
MPCM	$L = 9$, window size: 3×3, 5×5, 7×7
LIGP	k = 0.2, Local window size = 11×11
MGDWE	r = 2, Local window size = 7×7
NRAM	Patch size: 50×50, Slide step: 10, $\lambda = \frac{1}{\sqrt{\min(m,n)}}$
PSTNN	Patch size: 40×40, Slide step: 40, $\lambda = \frac{0.6}{\sqrt{\max(n_1,n_2)*n_3}}$, $\varepsilon = 1e^{-7}$

(1) Qualitative comparison. Figures 7–9 compare the detection results of the eight methods on three typical scenes of small infrared targets, where the detection methods are labeled in the top-left corner of each image. The target area is magnified in the lower-right corner to show the results of fine segmentation more visually. We used red, yellow, and green circles to indicate correctly detected targets, false positives, and missed detections, respectively.

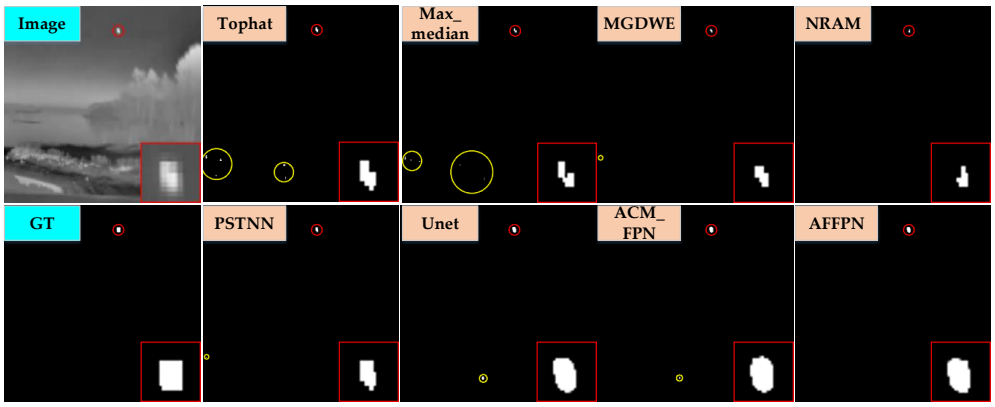


Figure 7. Qualitative results of the different methods for infrared scene 1.

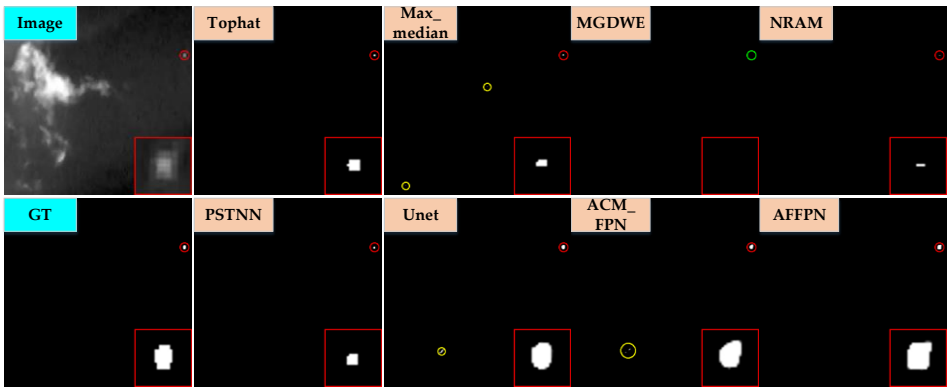


Figure 8. Qualitative results of the different methods for infrared scene 2.

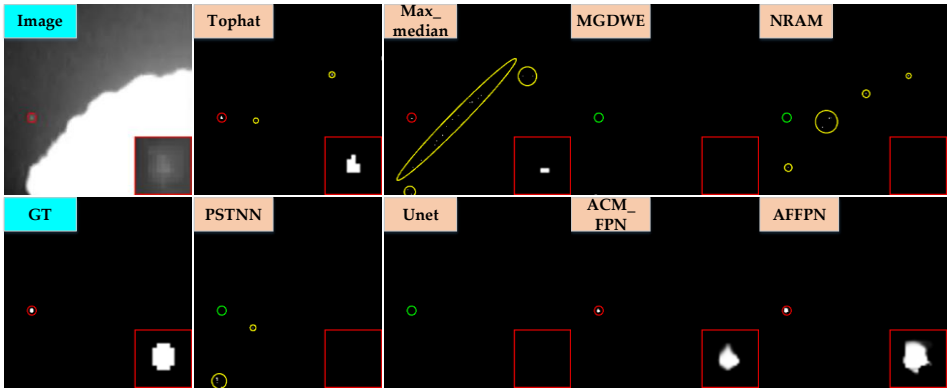


Figure 9. Qualitative results of the different methods for infrared scene 3.

The figures show that the filter-based top-hat and max-median methods are sensitive to noise and show varying numbers of false positives in different scenarios, indicating their strong response to background clutter and noise. The local rank-based methods yield more false positives and missed detections. Notably, these model-driven methods cannot fully segment the target shape accurately; they can only perceive the approximate position of the target. This is because traditional model-driven methods rely on hand-crafted features and a priori assumptions that do not adapt to the variation of various complex contexts in the SIRST dataset and are, therefore, less robust in different complex contexts. In the data-driven approach, U-Net does not consider the fusion of different feature layers and the association of global contextual information. Therefore, false positives occurred in scenes 1 and 2, and missed detections occurred in scene 3, further illustrating the importance of paying attention to the fusion between different feature layers and global contextual associations. ACM-FPN, which lacks a feature attention fusion module, yielded false positives in scenes 1 and 2. With the support of the AF module and ASPPM, AFFPN performs accurate localization and shape-contour detail segmentation of small infrared targets, achieving a better small target detection performance than the other methods.

Figures 10–12 show the three-dimensional visualization results of the eight methods for three typical small infrared target scenarios. Clearly, the top-hat and max-median methods have the most severe FPR, further illustrating their sensitivity to background clutter and noise. Other traditional methods (for example, NRAM and PSTNN) show severe false positives and missed detections in scenes with complex backgrounds. This is because the traditional methods largely depend on a priori assumptions and hyperparameter settings and do not adapt well to changes in the background complexity and target size. The CNN-based methods exhibit a better detection performance than traditional model-driven methods; however, U-Net inevitably experienced target loss in scene 4. U-Net and ACM-FPN output false positives in scene 5. With the attention-fusion module and ASPPM, which focuses on contextual information, the proposed method achieves more accurate detection of small infrared targets and adapts to the challenges of various complex backgrounds, and variations in the target size, resulting in a better performance.

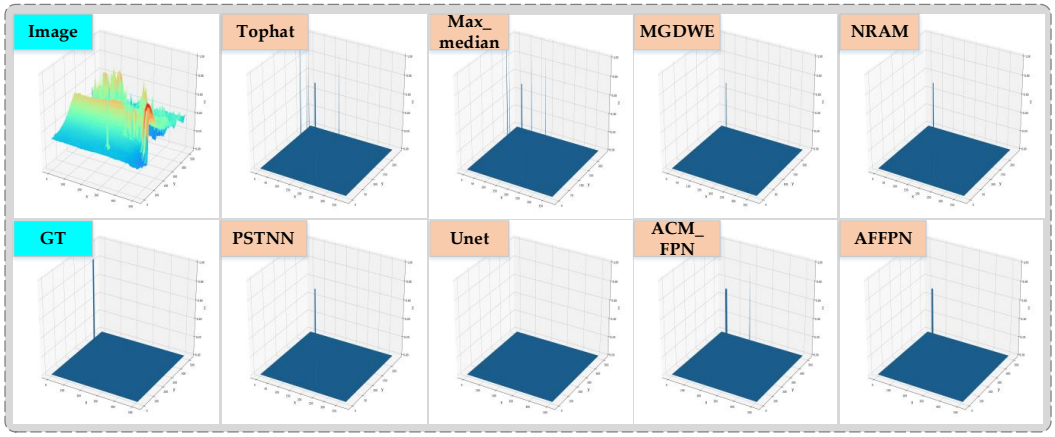


Figure 10. Three-dimensional representation of the results of the different methods for infrared scene 4.

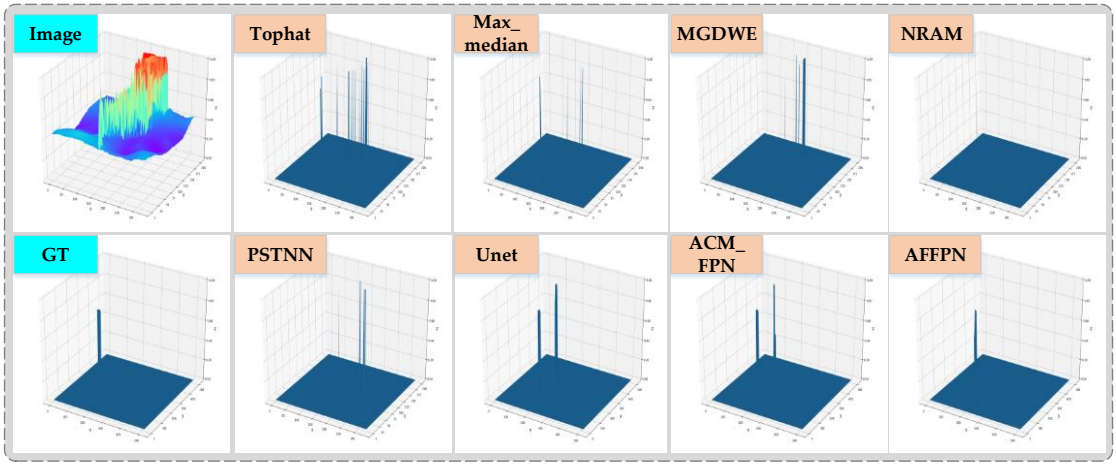


Figure 11. Three-dimensional representation of the results of the different methods for infrared scene 5.

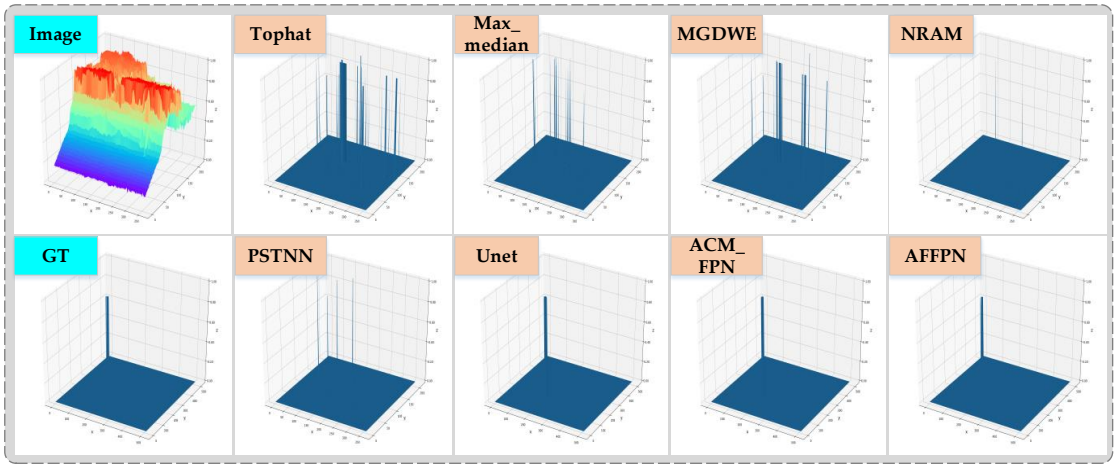


Figure 12. Three-dimensional representation of the results of the different methods for infrared scene 6.

(2) Numerical quantitative comparison. We obtained the predicted values of all the traditional model-driven methods, after which we eliminated low response regions by setting adaptive thresholds to suppress noise, calculated as follows:

$$T_{adaptive} = 0.5 \times max\_value(G) + 0.5 \times avg\_value(G) \quad (14)$$

where $max\_value(G)$ and $avg\_value(G)$ denote the maximum and average values of the output, respectively. All data-driven methods used the experimental parameters of the original authors.

Table 6 details the mIoU, nIoU, F-measure, AP, and AUC evaluation metrics for the 14 different methods. The best results in each column are highlighted in boldface red font and the second-best results are in boldface blue font. As can be seen in the table, AFFPN achieved the best performance on all the evaluation metrics. The significant increase in these evaluation metrics indicates that the proposed algorithm provides a significant improvement in both small infrared target detection and segmentation. The advantages of CNN-based methods over traditional model-driven approaches are clear. This is because the SIRST dataset contains several challenging images with complex background clutter, and different sizes and shapes of the target. The model-driven methods, which are based on the assumption of prior knowledge constraints, suppress the target while suppressing the background, and the hand-selected parameters limit the general performance of these methods. Consequently, model-driven methods detect the target but have difficulty achieving complete segmentation of the target.

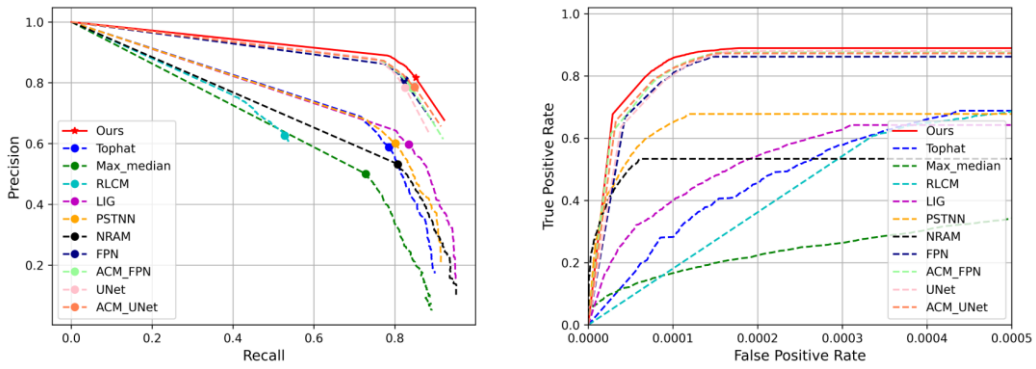
Table 6. Quantitative comparison with state-of-the-art methods on the SIRST dataset.

Methods	mIoU ($\times 10^{-2}$)	nIoU ($\times 10^{-2}$)	F-Measure ($\times 10^{-2}$)	AP ($\times 10^{-2}$)	AUC ($\times 10^{-2}$)
Top-hat	28.75	42.95	69.29	58.49	84.40
Max-median	15.65	25.43	62.40	41.50	74.96
RLCM	28.56	34.44	46.94	39.95	87.97
MPCM	21.35	24.54	65.98	39.73	72.65
LIGP	31.01	40.62	72.56	58.83	82.11
MGDWE	16.22	23.06	50.60	20.30	61.85
NRAM	24.99	32.39	67.82	48.38	76.69
PSTNN	39.68	48.16	71.73	59.31	83.89
FPN	72.18	70.41	80.39	75.90	93.10
U-Net	73.64	72.35	80.81	76.11	**94.01**
TBC-Net	73.40	71.30	—	—	—
ACM-FPN	73.65	72.22	81.60	**78.33**	93.79
ACM-U-Net	**74.45**	**72.70**	**81.68**	78.08	93.63
AFFPN(Ours)	**78.14**	**75.91**	**83.53**	**80.61**	**94.52**

The improvements achieved by AFFPN over the data-driven CNN-based approach are evident. This is due to the redesign of the backbone network so that it is tailored for small infrared target detection. First, the network constructs effective hierarchical global prior in the small infrared target feature extraction phase to reduce the loss of contextual information in the deep features and to better capture the global contextual prior information of small targets. Subsequently, we selectively enhance the semantic information and spatial location detail information of deep and shallow features of the CNN using the designed attention fusion module to focus on the contributions of different network layers to the feature representation of the small infrared target and achieve progressive feature fusion of different network layers through the multilayer feature fusion module. With these targeted network structures and module designs, the inherent features of the small infrared targets are retained in the deeper layers of the network, enhancing the small target feature representation capability of the network, and thus significantly improving the detection performance.

The PR and ROC curve results for several different methods are compared in Figure 13. AFFPN outperforms existing CNN-based and traditional model-driven approaches in all metrics, which indicates its ability to significantly suppress the background and accurately

detect targets while segmenting them accurately. The PR curve results (Figure 13a) show that the proposed method achieves the best precision and recall rates, which implies that it is able to guarantee the overall target localization and detection accuracy in challenging scenarios where the targets vary in size, shape, and location. The ROC curve results (Figure 13b) show that the proposed method achieves the best performance, with probabilistic detection (PD) responding rapidly to changes in the FPR. However, the performance of the traditional model-driven approaches largely depends on a priori assumptions; they cannot adapt to changes in complex backgrounds, and they do not perform well with respect to both the PR and ROC curve evaluation metrics.



(a) PR curve comparison with different methods (b) ROC curve comparison with different methods

Figure 13. PR and ROC results of AFFPN and state-of-the-art methods.

(3) Comparison of the inference performance. The inference performance is key to the practical deployment and application of unmanned platforms. The NVIDIA Jetson AGX Xavier development board has been widely used in a variety of unmanned platforms because of its high-performance computing capabilities. We deployed AFFPN on a stationary high-performance computer platform to compare its inference performance with those of other methods. We also implemented it on the NVIDIA Jetson AGX Xavier development board to further advance the application of AFFPN in real-world scenarios.

For the baseline (model-driven and data-driven) approaches, the average inference time for a single image was measured using a CPU-based implementation. We calculated the inference speeds of AFFPN on a CPU, GPU, and AGX Xavier development board separately, and used C, G, and B to, respectively, differentiate them in the test environment, as described in Section 4.2. The NVIDIA Jetson AGX Xavier development board was used in 30 W power mode for the measurement. In the calculation of the inference time, only the time taken by the algorithm to process the image was considered, ignoring the time taken for data preparation. Table 7 presents the processing times for a single image frame for the different methods, and the average inference time for a single image for AFFPN on different computational resources.

Table 7. Inference times for different algorithms.

Methods	**Top-Hat**	**Max-Median**	**RLCM**	**MPCM**	**LIGP**	**MGDWE**	**NRAM**
Times (s)	0.006	0.007	6.850	0.347	0.877	1.670	0.971
Methods	TBC-Net	U-Net	ACM-FPN	ACM-U-Net	Ours (C)	Ours (G)	Ours (B)
Times (s)	0.049	0.144	0.067	0.156	0.218	0.008	0.059

The inference times for different algorithms reported in their original papers are also listed in Table 7. The inference times of the algorithms are slightly different from those reported in the original papers because of the different computing platforms. AFFPN shows

an excellent inference performance on the CPU, although it is slower than the top-hat and max-median algorithms. Algorithms such as RLCM, MPCM, and MGDWE are slower, and their detection accuracy is equally poor.

AFFPN runs at approximately the same speed as the U-Net and ACM-U-Net methods on the CPU and achieves a frame rate of 125 frames per second (fps) on the GPU, which is a satisfactory inference performance. Notably, the AFFPN deployed on the NVIDIA Jetson AGX Xavier development board achieves an inference performance comparable to that of ACM-FPN, which fully demonstrates the feasibility of deploying the proposed network for practical applications on unmanned aerial vehicle platforms.

We compared the performance of AFFPN on the NVIDIA Jetson AGX Xavier development board in different power modes to further demonstrate that the AFFPN can support embedded platforms deployed in multiple application scenarios. The NVIDIA Jetson AGX Xavier development board supports three power modes: 10, 15, and 30 W, to suit different applications. Considering that deep learning accelerators are faster when processing data in batches than in single frames, and that in embedded scenarios, the embedded development board may need to simultaneously process video images captured by multiple sensors, we tested the average processing time of a single-frame image at different batch sizes for AFFPN. The inference speed results for different power and different batch sizes are listed in Table 8. The proposed method achieves frame rates of up to 21, 22, and 44 fps for batch sizes of 16 in the 10, 15, and 30 W power modes, respectively, which demonstrates the suitability of AFFPN for real-time and efficient target detection tasks in a wide range of embedded scenarios.

Table 8. AFFPN inference times for different powers and batch sizes.

Time (s)		Batch Size					
		1	**2**	**4**	**8**	**16**	**32**
Power Mode (W)	10	0.1203	0.0815	0.0624	0.0521	0.0472	0.0479
	15	0.1190	0.0804	0.0612	0.0514	0.0459	0.0461
	30	0.0593	0.0391	0.0299	0.0249	0.0227	0.0236

In the experiments, we considered the detection of small infrared targets as a pixel-level semantic segmentation task and presented the details of the experimental implementation and the dataset and evaluation metrics. Detailed ablation experiments were then performed to demonstrate the effectiveness of the structural design and network modules. This was followed by a description of the model-driven and data-driven approaches in the comparison experiments, which qualitatively compared the detection performance of several state-of-the-art approaches visually. Furthermore, AFFPN was compared with state-of-the-art approaches using a quantitative evaluation. Finally, AFFPN was deployed on the NVIDIA Jetson AGX Xavier development board to validate the possibility of using the AFFPN algorithm for practical applications. The effectiveness and reliability of AFFPN in the detection of single frames of small infrared target images is evident from the visualization results and the quantitative evaluation.

5. Conclusions

In this study, we proposed a novel small infrared target detection method called the AFFPN, comprising feature extraction and feature fusion modules. In the feature extraction stage, ASPPM was used to reduce the loss of contextual information in the deeper layers of the network and to better exploit the global contextual prior information of small targets. To focus on small target features in different layers, the network was designed with CA and SA to focus on the semantic and spatial location information of the target in the deep and shallow layers, respectively. This improves the fusion and utilization of the inherent information of small targets while retaining and focusing on the small infrared target features. Finally, a multiscale feature fusion module was proposed to further

improve the utilization of features. We compared the proposed method with state-of-the-art methods and conducted extensive ablation studies to verify the effectiveness of the individual modules. Our proposed method achieved the best performance on the publicly available SIRST dataset and has the ability to achieve accurate detection in complex scenes. Additionally, we deployed AFFPN on the NVIDIA Jetson AGX Xavier development board to evaluate its performance in real-time detection tasks. Lightweight model deployment of this kind has considerable potential for applications in areas such as infrared detection and search systems in unmanned aerial vehicles.

However, some unresolved issues, such as the efficiency of the feature fusion methods, the accuracy of deep feature representation, and the speed and reliability of the deployment, deserve further research. In future studies, we will continue to explore the practical applications of lightweight high-performance models for small infrared target detection.

Author Contributions: Conceptualization, Z.Z. and X.T.; methodology, Z.Z. and X.T.; software, Z.Z. and S.S.; supervision, B.S. and J.W.; validation, X.T., J.W. and P.W.; formal analysis, Z.Z. and X.T.; data curation, X.T.; writing—original draft preparation, Z.Z. and X.T.; writing—review and editing, P.W. and R.G.; project administration, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Hunan Provincial Innovation Foundation for Postgraduate CX20210020, Project supported by Provincial Natural Science Foundation of Hunan 2020JJ5672.

Data Availability Statement: The data used for training and test set SIRST are available at: https://github.com/YimianDai/sirst, accessed on 10 April 2022.

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

References

1. Rawat, S.; Verma, S.K.; Kumar, Y. Review on recent development in infrared small target detection algorithms. *Procedia Comput. Sci.* **2020**, *167*, 2496–2505. [CrossRef]
2. Tom, V.T.; Peli, T.; Leung, M.; Bondaryk, J.E. Morphology-based algorithm for point target detection in infrared backgrounds. In Proceedings of the Signal and Data Processing of Small Targets, Orlando, FL, USA, 22 October 1993; pp. 2–11. [CrossRef]
3. Deshpande, S.D.; Er, M.H.; Venkateswarlu, R.; Chan, P. *Max-Mean and Max-Median Filters for Detection of Small Targets*; SPIE: Bellingham, WA, USA, 1999. [CrossRef]
4. Chen, C.; Li, H.; Wei, Y.; Xia, T.; Tang, Y. A Local Contrast Method for Small Infrared Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 574–581. [CrossRef]
5. 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.
6. 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.* **2020**, *17*, 1822–1826. [CrossRef]
7. Han, J.; Moradi, S.; Faramarzi, I.; Zhang, H.; Zhao, Q.; Zhang, X.; Li, N. Infrared Small Target Detection Based on the Weighted Strengthened Local Contrast Measure. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1670–1674. [CrossRef]
8. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A. Infrared Patch-Image Model for Small Target Detection in a Single Image. *IEEE Trans. Image Process.* **2013**, *22*, 4996–5009. [CrossRef] [PubMed]
9. Lin, Z.; Ganesh, A.; Wright, J.; Wu, L.; Chen, M.; Ma, Y. Fast Convex Optimization Algorithms for Exact Recovery of a Corrupted Low-Rank Matrix. Available online: https://hdl.handle.net/2142/74352 (accessed on 15 August 2009).
10. Wang, X.; Peng, Z.; Kong, D.; Zhang, P.; He, Y. Infrared dim target detection based on total variation regularization and principal component pursuit. *Image Vis. Comput.* **2017**, *63*, 1–9. [CrossRef]
11. 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]
12. Wang, H.; Zhou, L.; Wang, L. 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 2019, Seoul, Korea, 27 October–2 November 2019; pp. 8508–8517.
13. Gao, Z.; Dai, J.; Xie, C. Dim and small target detection based on feature mapping neural networks. *J. Vis. Commun. Image Represent.* **2019**, *62*, 206–216. [CrossRef]
14. McIntosh, B.; Venkataramanan, S.; Mahalanobis, A. Infrared Target Detection in Cluttered Environments by Maximization of a Target to Clutter Ratio (TCR) Metric Using a Convolutional Neural Network. *IEEE Trans. Aerosp. Electron. Syst.* **2021**, *57*, 485–496. [CrossRef]

15. Du, J.; Huanzhang, L.; Hu, M.; Zhang, L.; Xinglin, S. CNN-based infrared dim small target detection algorithm using target oriented shallo— deep features and effective small anchor. *Let Image Processing* **2020**, *15*, 1–15.
16. Zhao, B.; Wang, C.-P.; Fu, Q.; Han, Z.-S. A Novel Pattern for Infrared Small Target Detection with Generative Adversarial Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4481–4492. [CrossRef]
17. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric Contextual Modulation for Infrared Small Target Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 949–958.
18. 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. In *IEEE Transactions on Aerospace and Electronic Systems*; IEEE: New York, NY, USA, 2022.
19. Hou, Q.; Wang, Z.; Tan, F.-J.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust Infrared Small Target Detection Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 7000805. [CrossRef]
20. Hou, Q.; Zhang, L.; Tan, F.-J.; Xi, Y.; Zheng, H.; Li, N. ISTDU-Net: Infrared Small-Target Detection U-Net. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 7506205. [CrossRef]
21. Ma, T.; Yang, Z.; Wang, J.; Sun, S.; Ren, X.; Ahmad, U. Infrared Small Target Detection Network With Generate Label and Feature Mapping. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6505405. [CrossRef]
22. Wang, K.; Du, S.; Liu, C.; Cao, Z. Interior Attention-Aware Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5002013. [CrossRef]
23. Wang, A.; Li, W.; Wu, X.; Huang, Z.; Tao, R. MPANet: Multi-Patch Attention for Infrared Small Target object Detection. *arXiv* **2022**, arXiv:2206.02120.
24. Chen, Y.; Li, L.; Liu, X.; Su, X.; Chen, F. A Multi-task Framework for Infrared Small Target Detection and Segmentation. *arXiv* **2022**, arXiv:2206.06923.
25. Zhou, H.; Tian, C.; Zhang, Z.; Li, C.; Xie, Y.; Li, Z. PixelGame: Infrared small target segmentation as a Nash equilibrium. *arXiv* **2022**, arXiv:2205.13124.
26. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
27. 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. *arXiv* **2021**, arXiv:2106.00487.
28. Zhao, T.; Wu, X. Pyramid Feature Attention Network for Saliency Detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3080–3089.
29. 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]
30. Qin, Y.; Bruzzone, L.; Gao, C.; Li, B. Infrared Small Target Detection Based on Facet Kernel and Random Walker. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7104–7118. [CrossRef]
31. Liu, J.; He, Z.; Chen, Z.; Shao, L. Tiny and Dim Infrared Target Detection Based on Weighted Local Contrast. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1780–1784. [CrossRef]
32. Deng, Q.; Lu, H.; Tao, H.; Hu, M.; Zhao, F. Multi-Scale Convolutional Neural Networks for Space Infrared Point Objects Discrimination. *IEEE Access* **2019**, *7*, 28113–28123. [CrossRef]
33. Shi, M.; Wang, H. Infrared Dim and Small Target Detection Based on Denoising Autoencoder Network. *Mob. Netw. Appl.* **2020**, *25*, 1469–1483. [CrossRef]
34. Wang, K.; Li, S.; Niu, S.; Zhang, K. Detection of Infrared Small Targets Using Feature Fusion Convolutional Network. *IEEE Access* **2019**, *7*, 146081–146092. [CrossRef]
35. 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* **2020**, arXiv:2001.05852.
36. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
37. Zhang, T.; Cao, S.; Pu, T.; Peng, Z. AGPCNet: Attention-Guided Pyramid Context Networks for Infrared Small Target Detection. *arXiv* **2021**, arXiv:2111.03580.
38. Zhang, H.; Goodfellow, I.J.; Metaxas, D.N.; Odena, A. Self-Attention Generative Adversarial Networks. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York City, NY, USA, 2019.
39. Liu, H.; Chen, T.; Guo, P.; Shen, Q.; Cao, X.; Wang, Y.; Ma, Z. Non-local Attention Optimized Deep Image Compression. *arXiv* **2019**, arXiv:1904.09757.
40. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.-S. CBAM: Convolutional Block Attention Module. In Proceedings of the European conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
41. Fu, J.; Liu, J.; Tian, H.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149.
42. Park, J.; Woo, S.; Lee, J.-Y.; Kweon, I.-S. BAM: Bottleneck Attention Module. *arXiv* **2018**, arXiv:1807.06514.
43. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11531–11539.

44. Zhang, H.; Zu, K.; Lu, J.; Zou, Y.; Meng, D. EPSANet: An Efficient Pyramid Split Attention Block on Convolutional Neural Network. *arXiv* **2021**, arXiv:2105.14447.
45. Zhang, Q.-L.; Yang, Y. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239.
46. Lin, T.-Y.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
47. Zhang, Y.; Hsieh, J.-W.; Lee, C.-C.; Fan, K.-C. SFPN: Synthetic FPN for Object Detection. *arXiv* **2022**, arXiv:2203.02445.
48. Gong, Y.; Yu, X.; Ding, Y.; Peng, X.; Zhao, J.; Han, Z. Effective Fusion Factor in FPN for Tiny Object Detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 1159–1167.
49. Tong, X.; Sun, B.; Wei, J.; Zuo, Z.; Su, S. EAAU-Net: Enhanced Asymmetric Attention U-Net for Infrared Small Target Detection. *Remote. Sens.* **2021**, *13*, 3200. [CrossRef]
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
51. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890. [CrossRef]
52. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
53. Fang, H.; Xia, M.; Zhou, G.; Chang, Y.; Yan, L. Infrared Small UAV Target Detection Based on Residual Image Prediction via Global and Local Dilated Residual Networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 7002305. [CrossRef]
54. Duchi, J.C.; Hazan, E.; Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J. Mach. Learn. Res.* **2011**, *12*, 2121–2159.
55. Rahman, M.A.; Wang, Y. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. In *Advances in Visual Computing, Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2016.
56. Deng, H.; Sun, X.; Liu, M.; Ye, C.; Zhou, X. Small Infrared Target Detection Based on Weighted Local Difference Measure. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4204–4214. [CrossRef]
57. Zhang, H.; Zhang, L.; Yuan, D.; Chen, H. Infrared small target detection based on local intensity and gradient properties. *Infrared Phys. Technol.* **2018**, *89*, 88–96. [CrossRef]
58. Deng, H.; Sun, X.; Liu, M.; Ye, C.; Zhou, X. Infrared small-target detection using multiscale gray difference weighted image entropy. *IEEE Trans. Aerosp. Electron. Syst.* **2016**, *52*, 60–72. [CrossRef]
59. Zhang, L.; Peng, L.; Zhang, T.; Cao, S.; Peng, Z. Infrared Small Target Detection via Non-Convex Rank Approximation Minimization Joint l2, 1 Norm. *Remote. Sens.* **2018**, *10*, 1821. [CrossRef]
60. Zhang, L.; Peng, Z. Infrared Small Target Detection Based on Partial Sum of the Tensor Nuclear Norm. *Remote. Sens.* **2019**, *11*, 382. [CrossRef]