remote sensing

MDPI

Article

Novel Asymmetric Pyramid Aggregation Network for Infrared Dim and Small Target Detection

Guangrui Lv, Lili Dong *, Junke Liang and Wenhai Xu

Schoolof Information Science and Technology, Dalian Maritime University, Dalian 116026, China
* Correspondence: donglili@dlmu.edu.cn

Abstract: Robust and efficient detection of small infrared target is a critical and challenging task in infrared search and tracking applications. The size of the small infrared targets is relatively tiny compared to the ordinary targets, and the sizes and appearances of the these targets in different scenarios are quite different. Besides, these targets are easily submerged in various background noise. To tackle the aforementioned challenges, a novel asymmetric pyramid aggregation network (APANet) is proposed. Specifically, a pyramid structure integrating dual attention and dense connection is firstly constructed, which can not only generate attention-refined multi-scale features in different layers, but also preserve the primitive features of infrared small targets among multi-scale features. Then, the adjacent cross-scale features in these multi-scale information are sequentially modulated through pair-wise asymmetric combination. This mutual dynamic modulation can continuously exchange heterogeneous cross-scale information along the layer-wise aggregation path until an inverted pyramid is generated. In this way, the semantic features of lower-level network are enriched by incorporating local focus from higher-level network while the detail features of high-level network are refined by embedding point-wise focus from lower-level network, which can highlight small target features and suppress background interference. Subsequently, recursive asymmetric fusion is designed to further dynamically modulate and aggregate high resolution features of different layers in the inverted pyramid, which can also enhance the local high response of small target. Finally, a series of comparative experiments are conducted on two public datasets, and the experimental results show that the APANet can more accurately detect small targets compared to some state-of-the-art methods.

Keywords: infrared small target; dual attention; dense connection; pair-wise asymmetric combination; inverted pyramid; recursive asymmetric fusion

Citation: Lv, G.; Dong, L.; Liang, J.; Xu, W. Novel Asymmetric Pyramid Aggregation Network for Infrared Dim and Small Target Detection. *Remote Sens.* **2022**, *14*, 5643. https://doi.org/10.3390/rs14225643

Academic Editor: Andrzej Stateczny

Received: 6 September 2022
Accepted: 1 November 2022
Published: 8 November 2022

1. Introduction

In the past few years, infrared small target detection has been widely applied in various fields such as remote sensing, medical imaging, early warning systems, and maritime surveillance [1,2]. However, infrared small objects often lack sufficient texture and shape information due to the long imaging distance. It is difficult to extract the small infrared target features effectively because there are fewer available target pixels and the background occupies most of the pixels. In addition, small objects usually have weak contrast compared to the background in complex imaging environmental conditions. In these situations, small objects are easily swamped by heavy noise and clutter background (as shown in Figure 1a). Moreover, different radiation, inherent sensor noise and natural factors can also affect the infrared imaging quality. More seriously, the appearance shape and size of small infrared targets in diverse scenarios are quite different, which will further reduce the stability of small target detection (as shown in Figure 1b). All in all, the aforementioned characteristics make the robust and precise small infrared target detection a complex and compelling task.

Traditional methods rely on different assumptions to design some handcrafted features. Specially, some methods based on background estimation [3,4], local contrast metrics [5–8], and non-local auto-correlation properties [9–11] were proposed to detect the small infrared

targets. However, these conventional methods are constrained by various assumptions and empirical knowledge, and cannot achieve good generalization, especially for those situations with complex and diverse backgrounds.

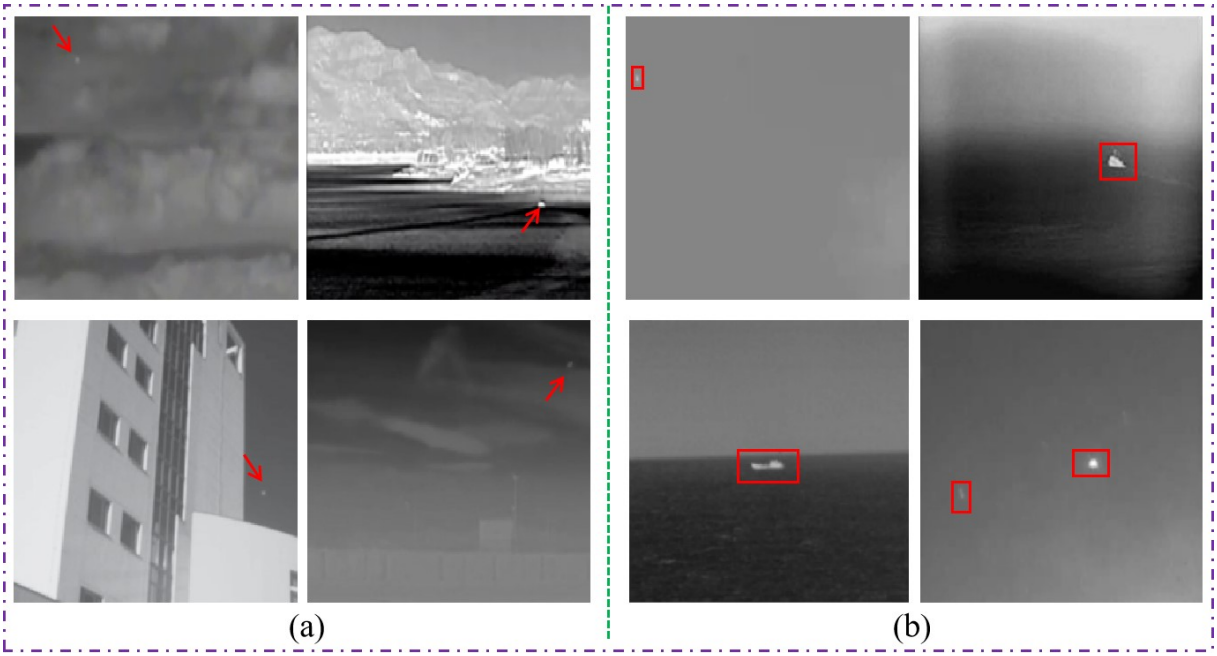


Figure 1. Examples of small infrared targets. (**a**) Small targets submerged in complex environments, as indicated by the red arrows. (**b**) Small targets of different scales, as indicated by the red boxes.

At present, convolutional neural networks (CNN) have well overcome the limitations of traditional handcrafted feature extraction. In particular, the widespread application of fully convolutional network (FCN) [12] has significantly promoted the development of target segmentation and detection tasks. The FCN can increase the receptive field and capture discriminative features through continuous down-sampling operations, but it results in a significant reduction in spatial resolution and thus losing fine image detail features. Several studies are working to address the above issues. E.g., refs. [13,14] adopt the "Unet" of encoder-decoder architecture to capture and preserve more information from low-level spatial features. Some strong pyramid feature architectures (e.g., pyramid pooling module (PPM) [15] and atrous spatial pyramid pooling (ASPP) [16–18]) are built to enlarge the receptive field of high-level network features and thus enrich the feature representation of different objects. These existing networks can express features well, but it is infeasible to directly utilize these deep CNN-based methods to detect and segment small infrared targets! Usually, small infrared targets vary widely in size, ranging from point targets covering only one pixel to extended targets containing tens of pixels, that is, their sizes are small, but they still have different sizes and shapes. Furthermore, the small targets occupies only a tiny component of the overall infrared image, and these targets are often easily confused in the messy and various background.

To address the problem of small targets with different scales and their appearances similar to the background noise, the local detail features and the multi-scale difference features need to be mined. For example, pyramid contextual attention [19], multiple contextual attention [20] and local similarity pyramid [21] have been designed to highlight the local features. However, these designed modules are only applied to the feature maps of the top-level network, ignoring the effectiveness of multi-level features. In order to fully mine multi-level features, most methods only use simple merging operation [21,22]. But, these merging operations cannot explore the relationship of multi-layer and cross-scale features to achieve true complementary connections among these features [23]. In general, the coarse information (e.g., lines, edges, corners, etc.) in the low-level network features

is more diverse, while the abstract information about the target in the high-level network features is richer. And the difference between network characteristics of different levels is usually considered as a semantic gap [23]. Therefore, merging feature simply will result in the newly generated multi-scale fusion features to still retain rough background information, which affects the accuracy of small infrared object detection. Recently, some methods of asymmetric features modulation [24,25] can incorporate cross-layer features in a gated manner to detect small infrared target, however, their methods ignore the dynamic modulation between local semantics and local details.

Based on the above discussions and the aforementioned limitations, a novel asymmetric pyramid aggregation network (APANet) is proposed. Specifically, inspired by some multi-scale feature learning methods [13,15,19,21], we integrate dual attention into different stages of the network to enhance spatial features on different scales, and then build dense connections to continuously preserved the enhanced detail information of small infrared targets in the multi-scale features, especially in deep-level features. Moreover, different from existing single-level asymmetric features modulation methods [24,25], we construct the multi-level and multi-path asymmetric local modulation to interact with higher-level local semantic and lower-level fine details dynamically and sequentially. In detail, the adjacent information in multi-scale enhancement features are first aggregated through pair-wise asymmetric modulation to generate the multi-level inverted pyramid. Among them, the aggregated features in the inverted pyramid can contain heterogeneous information of small targets from two and more adjacent scales gradually. Then, the recursive asymmetric modulation is designed to highlight and preserve high-response detail cues of small targets at different levels of the inverted pyramid. Overall, the consistency of details and semantics of small targets can be enhanced adaptively under these different gating aggregation paths.

The main contributions of our work are as follows:

1. An end-to-end gated multiple pyramid structure is proposed for detecting small infrared targets. Specially, the pyramid structure is first built to encode multi-scale enhanced features of small infrared targets. And then the inverted pyramid structure is built to decode asymmetric local information of small infrared targets in multi-scale and multi-level features.
2. A densely connected feature pyramid extraction module is proposed to continuously enhance and retain the details of small infrared target in different scale features. Specially, based on the different forms of information flow transmission on the backbone network, two different variants of feature pyramid extraction are designed, which can transfer detailed features enhanced by dual attention of small target from lower-level large-scale space to higher-level small-scale space.
3. An enhanced asymmetric feature pyramid aggregation module is proposed to dynamically highlight the fine details of small targets and suppress complex backgrounds. The module can modulate and aggregate cross-layer local information in pairwise asymmetric manner and recursive asymmetric manner, respectively. In particular, two different aggregation paths, each with two different interaction strategies: parallel gated fusion and hierarchical gated fusion.

2. Related Work

Our proposed small infrared target detection network mainly involves the following several aspects.

2.1. Small Infrared Target Detection

Conventional methods design filters or modules based on a priori knowledge. Some earlier methods based on background estimation (e.g., top-hat morphological filter [3] and max-mean/max-median filter [26]) were proposed to detect targets by subtracting the calculated background from the infrared image. Li et al. [27] constructed a combination of directional morphological filter, multi-directional improved top-hat filter, and histogram of oriented morphological filter to detect real small objects and eliminate false alarms.

Obviously, these methods cannot stably adapt to those scenarios where the target size varies greatly, because different parameters based on morphological structure need to be specially designed and continuously adjusted for different scenarios.

Some other methods based on local contrastive saliency usually take the difference between the central pixel and surrounding pixels in the fixed-size local patch as the ratio of local contrast, and traverse the entire image to measure the local contrast of the image. For example, both local contrast metric (LCM) [5] and improved local contrast measure (ILCM) [6] can capture the pixel-contrast of local patches by designing special local filters. Furthermore, multiscale patch-based contrast measure (MPCM) [28] and relative local contrast measure (RLCM) [29] are successively proposed. Both MPCM and RLCM can calculate the local dissimilarity of multi-scale image patches. Compared to MPCM, RLCM adds the feature representation of the internal intensity about the object. Zhang et al. [7] also explored the local intensity and gradient of small targets to suppress clutter and enhance target features. Li et al. [8] first extracted some candidate targets, and then constructed the local contour contrast descriptor to identify true infrared target. In general, small objects tend to be those pixels with significant local contrast. However, these methods are obviously not suitable for the situation where the target is close to the background.

Some methods based on non-local auto-correlation properties assume that the object and background have a low-rank and sparse relationship, so the task of small object detection is approximately transformed into the operation of low-rank sparse matrix factorization. For example, these methods of low-rank based infrared patch-image (IPI) model [9], patch image model with local and global analysis (PILGA) [10], and partial sum of tensor nuclear norm (PSTNN) [11] joint weighted l_1 norm utilize the non-local self-correlation property to suppresses background and preserves target. Whereas, the detection methods that combine the background and target characteristics require large amount of computation.

In summary, traditional methods rely heavily on certain assumption and prior knowledge, which makes them lack generalization ability beyond prior knowledge when detecting infrared small targets. That is, traditional methods are suitable for specific application fields, and are limited to other complex application backgrounds.

In recent years, CNN-based methods have been applied to small infrared target detection. Earlier, Liu et al. [30] designed a 5-layer multi-layer perceptron (MLP) network for extracting infrared targets. Subsequently, Shi et al. [31] converted the small target detection task into a noise removal problem, and then combined CNN and denoising autoencoder for detecting small infrared target. Zhao et al. [32] designed a U-Net structure combined a semantic constraint mechanism for small infrared target detection. Dai et al. [24] utilized a bidirectional path with global attention modulation and point-wise attention modulation to retain more feature information for small infrared target detection. Huang et al. [21] progressively aggregate local similar pyramid features on the top layer network into low-level features at different scales sequentially to improve the performance of small infrared target detection.

Compared to traditional methods, CNN-based methods can automatically learn the features of small targets adaptively, and can obtain better detection effect significantly. However, the effect of these methods is still limited because these approaches are less robust against scenarios such as changeable small targets, dim small targets, and complex backgrounds.

2.2. Pyramid Structure

Some methods have utilized multi-scale pyramid structures to obtain dense receptive fields. For example, feature pyramid network (FPN) [33] builds a multi-scale pyramid hierarchical structure based on deep CNN, which upsamples high-level features and performs cross-level connections with lower-level features from top to down. In order to simultaneously utilize the the discriminative semantics of high-level network and high-resolution features of low-level network, it predicts each layer of features to improve

the object detection. In addition, PSPNet [15] performs a spatial PPM on the multi-layer down-sampled feature maps to capture multi-scale local and global information, and these local and global cues are combined to make the final prediction more reliable. Unlike PSPNet [15], adaptive pyramidal context network (APCNet) [19] first learns multi-scale contextual representations adaptively and then stacks these different scales of contextual information in parallel like the operation of PPM. DeeplabV2 [16] constructs an ASPP module, which adopts parallel atrous convolutional layers with different ratios to captures multi-scale information based on multiple parallel convolutional layers with different atrous ratios, and then fuse the local and global information by concatenating the different scales features to fuse the local and global context. Subsequently, DeepLabV3 [17] adds a global pooling branch to improve the ASPP module based on the parallel structure in DeepLabv2. Compared with DeepLabv3 [17], DeepLabv3+ [18] uses the entire network of DeepLabV3 as an encoder to extract features at different scales, and introduces a decoder module, which fuses features of different layers to improve the detection of object boundary. DenseASPP [34] constructs several paths to connect a series of atrous convolutions in a dense manner, which can efficiently generate spatial features covers a larger scale range.

2.3. Attention Mechanism

The attention mechanism can discover important content and give it more focus. As a way of adaptive learning, attention mechanism has been widely designed and applied in the related research of deep network. For example, squeeze-and-excitation network (SENet) [35] capture the global correlation among channels to enhances informative feature maps. Different from SENet, convolutional block attention module (CBAM) [36] computes and infers different attention maps along channel and spatial dimensions sequentially for refining features of different dimensions. While selective kernel network (SKNet) [37] selects the size of the receptive field dynamically based on different convolution kernel weights. And spatial gated attention (SGA) [38] structure generates a gated attention mask to suppress background clutter while focus on regions of interest. In addition, there has been some research on attention mechanisms to explore the spatial dependencies of pixels. For example, non-local module (NLM) [39] constructs self-attention mechanism to captures the contextual dependencies among different pixels in a single spatial map. And dual attention network (DANet) [20] learns long-range semantic dependencies of both spatial and channel dimensions by designing spatial and channel attention, respectively.

2.4. Cross-Layer Feature Aggregation

How to better aggregate cross-layer features of deep network has always been a task worthy of research. So far, some research on multi-layer feature fusion has been achieved. For example, Both U-Net [13] and SegNet [14] combine coarse features from lower network layers and rich semantic features from higher network layers hierarchically. Li et al. [40] concatenated features from different layers directly to enrich feature representation. Zhang et al. [41] transformed different layers of features into several different resolutions, and these features were then used to output the final prediction result at a specific resolution. Recently, Li et al. [42] designed the feature pyramid attention model, which can capture high-level modulation information based on global channel attention [35] to guide lower-level features in skip connections. Dai et al. [24] proposed asymmetric contextual modulation (ACM) network based on global attention and point-wise attention to exchange shallow subtle details and deep rich semantics to detect small infrared targets. Huang et al. [21] constructed multi-scale feature fusion to aggregate local similar pyramid fusion features at the top-level network into low-level features at different scales progressively to improve infrared small target detection. Zhang et al. [25] proposed attention guided pyramidal contextual structure, which focuses on exploring the contextual relationships of top-level features and cross-layer asymmetric feature modulation(AFM) to improve the small infrared target detection.

3. Method

The APANet method comprises of two principal parts: a densely connected feature pyramid extraction module (as shown in Figure 2a) and an enhanced asymmetric feature pyramid aggregation module (as shown in Figure 2b). Specially, multi-scale pyramid feature extractor based on channel-spatial dual attention (CSDA) modulation is constructed, and the detailed features in the shallow network are densely transferred to the deep network to maintain the detailed features of small infrared objects in the multi-scale space. Moreover, adjacent cross-scale features from the pyramid extraction module are first mined by pair-wise asymmetric combination (PWAC) to obtain consistency between spatially finer shallow features and semantically richer deep features of small infrared objects. The PWAC is performed layer-wise to generate multi-scale aggregated contextual information until an inverted pyramid is built. Then, a recursive asymmetric fusion (RAF) mechanism is constructed to further learn the highly responsive target features among cross-level local contextual interaction of the same scale in the inverted pyramid for the detection of small infrared objects. In the following, the specific details of the different components in our proposed APANet will be introduced.

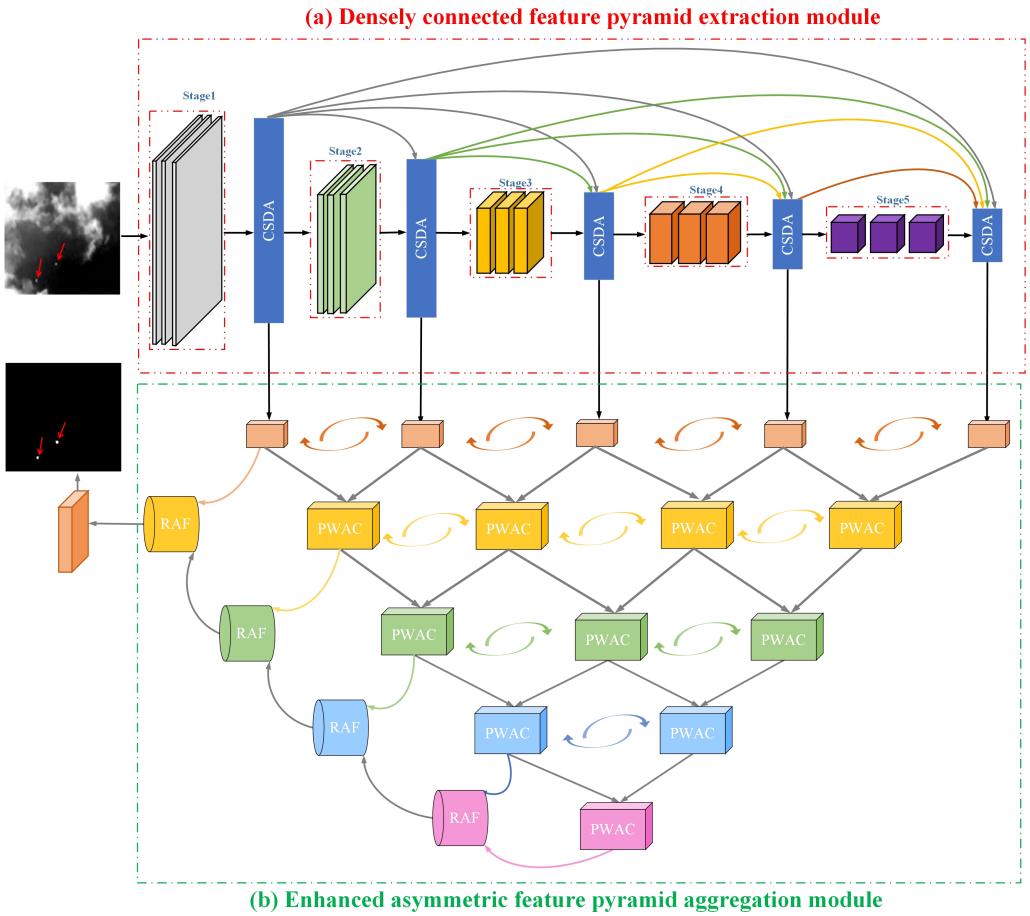


Figure 2. An illustration of the proposed novel asymmetric pyramid aggregation network (APANet). (**a**) Densely connected feature pyramid extraction module. Input images are first fed into the feature pyramid extraction module to extract multi-scale features. Note that, features from different scales are adaptively enhanced by a channel-spatial dual attention (CSDA), and enhanced features from the shallow large-scale space are intensively transferred to the deep small-scale space. (**b**) Enhanced asymmetric feature pyramid aggregation module. The features of adjacent scales are interacted bidirectionally through pair-wise asymmetric combination (PWAC), and these features are aggregated layer-wise in the way of an inverted pyramid. Then, recurrent asymmetric fusion (RAF) is exploited to successively integrate the leftmost multi-level features in the inverted pyramid.

3.1. Densely Connected Feature Pyramid Extraction Module

With the increase of spatial pooling operations in the network layer, these dim and small targets are easily lost in deep-level networks. Therefore, we should construct a densely connected feature pyramid extraction module to extract multi-level features of infrared dim and small targets and maintain the features of these targets in deeper network layers.

In general, the resolution of the feature map is high and the detail features are clearer in the shallow network, while the resolution of the feature map is low and the semantic information is richer in deeper network. Inspired by the DenseNet [43], we try to transfer the information of shallow features in the network to the deep features of different scales one by one. However, unlike DenseNet which uses dense skip connections to bridge features, we design CSDA to adaptively enhance features at different scales when bridging features, as shown in Figure 2a.

Specifically, the $CSDA(\cdot)$ comprises of two attention units connected in series. Assuming the spatial feature map in the feature pyramid is $s_i \in R^{C\times H\times W}$, where C represents the channel of s_i, H represents the height of s_i, and W represents the width of s_i, respectively. Then, s_i is sequentially processed by the 1D attention map $Z^c(s_i) \in R^{C\times 1\times 1}$ in channel dimension and the 2D attention map $Z^s(s_i) \in R^{1\times H\times W}$ in spatial dimension, as shown in Figure 3. The channel attention interaction can be communicated as follows:

$$Z^c(s_i) = \sigma\left[C2D_1(P_{\max}(s_i)) + C2D_1(P_{avg}(s_i))\right] \tag{1}$$

$$s_i^c = Z^c(s_i) \otimes s_i \tag{2}$$

where σ represents the *sigmoid* function, $C2D_1(\cdot)$ represents the shared convolution operation of the convolution kernel 1×1, $P_{avg}(\cdot)$ and $P_{\max}(\cdot)$ denote the global average pooling and global maximum pooling, respectively. The $Z^c(s_i)$ is multiplied element-wise with s_i to generate the channel attention-enhanced features s_i^c.

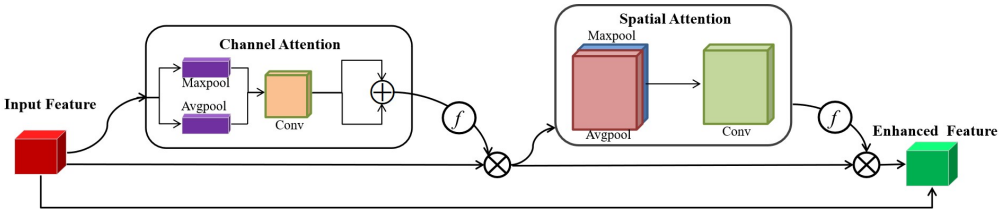


Figure 3. The illustration of the channel and spatial dual attention module.

Like the channel attention calculation process, the spatial attention calculation can be summed up as:

$$Z^s(s_i) = \sigma\left[C2D_7(f^c_{\max}(s_i^c), f^c_{avg}(s_i^c))\right] \tag{3}$$

$$s_i^p = Z^s(s_i) \otimes s_i^c \tag{4}$$

where $C2D_7(\cdot)$ rerepresents the convolution operation of the convolution kernel 7×7, $f^c_{\max}(\cdot)$ denotes the maximum pooling, $f^c_{avg}(\cdot)$ denotes average pooling. The $Z^s(s_i)$ is multiplied element-wise with s_i^c to generate the dual attention-enhanced features s_i^p.

Then, a skip connection is introduced to add s_i to s_i^p, which can preserve the information of the original input features. So far, the multi-dimensional refinement features enhanced by $CSDA(\cdot)$ are obained.

As shown in Figure 2a, the CSDA is embed into different stages of the backbone network to enhance multi-scale features. Then, to enhance deep propagation ability of spatial fine details of small infrared targets, two different densely connected mechanisms (as shown in Figure 4), densely connected multi-scale feature (DCMSF) and residual-based densely connected multi-scale feature (rb-DCMSF), are designed according to the transmission mode of information flow on the backbone network. Obviously, these two variants can fully retain the spatial details generated by shallow level network in the deep

level semantic features of the network. However, their difference lies in the spread of information flow on the backbone network.

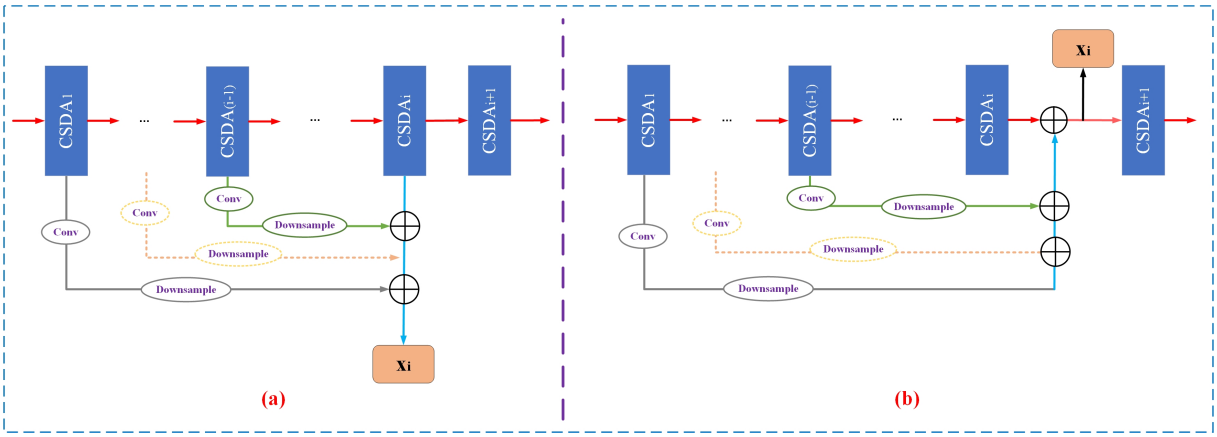


Figure 4. Densely connected feature pyramid extraction module. (**a**) Densely connected multi-scale features. (**b**) Resisual-based densely connected multi-scale features.

(1) DCMSF

As shown in Figure 4a, the feature pyramid extraction process of DCMSF is as follows:

$$s_i = \begin{cases} f_{c_3}^3(x_0), & \text{if } i = 1 \\ f_{c_3}^3(CSDA(s_{i-1})), & 2 \le i \le 5 \end{cases} \tag{5}$$

where x_0 denotes the input image, $f_{c_3}^3(\cdot)$ represents the 3×3 convolution block of different stage, and s_i denote the multi-scale feature generated from different stage.

(2) rb-DCMSF

As shown in Figure 4b, the feature pyramid extraction process of rb-DCMSF is as follows:

$$s_i = \begin{cases} f_{c_3}^3(x_0), & \text{if } i = 1 \\ f_{c_3}^3(CSDA(s_{i-1}) + \sum_{j=1}^{i-1} F_d(F_{c_1}(s_j))), & 2 \le i \le 5 \end{cases} \tag{6}$$

where $F_{c_1}(\cdot)$ represents the 1×1 convolution operation, and $F_d(\cdot)$ represents the down-sampling operation.

Then, based on DCMSF or rb-DCMSF, the densely connected multi-scale feature can be generated as follows:

$$x_i = \begin{cases} CSDA(s_i), & \text{if } i = 1 \\ CSDA(s_i) + \sum_{j=1}^{i-1} F_d(F_{c_1}(x_j)), & 2 \le i \le 5 \end{cases} \tag{7}$$

Based on the operations of Equation (7), spatial features $\{x_i\}_{i=1,\ldots,5}$ of five different scales in the densely connected feature pyramid extraction module can be obtained.

3.2. Enhanced Asymmetric Feature Pyramid Aggregation Module

The existing asymmetric feature fusion [24,25] methods can modulate higher-level global semantics and lower-level details for small infrared target detection, but they do not pay attention to the importance of high-level local semantics, nor to the role of intensive cross-layer modulation in feature fusion. Inspired by multi-scale pyramid feature extraction [21], an enhanced asymmetric feature pyramid aggregation module is proposed to intensively modulate cross-layer local information to highlight the characteristics of small infrared targets, as shown in Figure 2b. It mainly includes two different asymmetric modu-

lation mechanisms, namely PWAC and RAF. In the following, the enhanced asymmetric feature pyramid aggregation module will be elaborated.

3.2.1. Pair-Wise Asymmetric Combination

Generally, both semantic features from higher-layer networks and detail features from lower-layer networks are very important for small infrared target detection. Therefore, it is worth studying how to better preserve the inherent characteristics of the original spatial features and better match different spatial features in cross-layer feature fusion. In particular, we design two variants of PWAC from different perspectives, namely parallel asymmetric combination (PAC) and hierarchical asymmetric combination (HAC), as shown in Figure 5. Among them, PAC can more retain the inherent characteristics of the original features in cross-layer fusion, while HAC can more emphasize the importance of feature matching in cross-layer fusion.

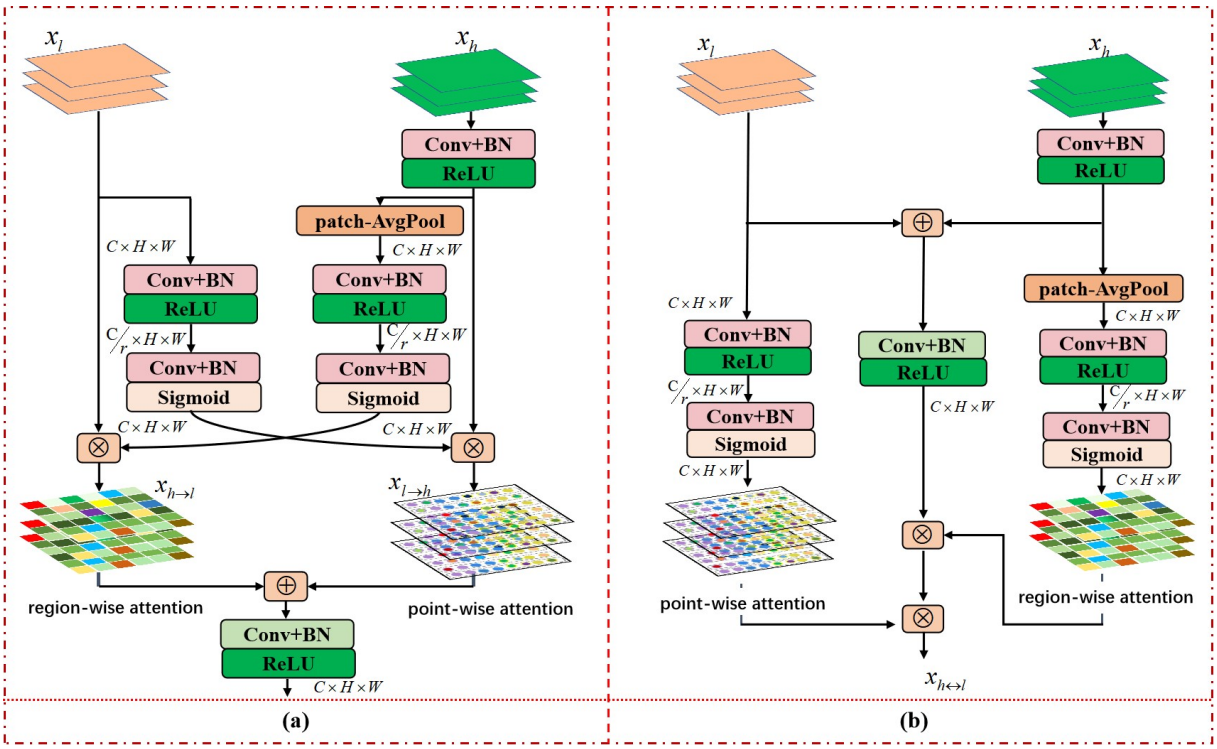


Figure 5. The illustration of the pair-wise asymmetric combination. (**a**) Parallel asymmetric combination. (**b**) Hierarchical asymmetric combination.

(1) Parallel asymmetric combination

The ACM [24] designs top-down attentional modulation with global average pooling (GAP) and bottom-up attentional modulation with point-wise convolution (PWConv) to exchange semantic information and spatial details in a parallel asymmetric manner. However, this top-down global channel context signal is not necessarily suitable for small infrared targets. With the increasing number of network layers, dim and small targets are easily overwhelmed by the background on the high-level features and their features are greatly weakened in the GAP. Therefore, in our work, local region context should be exploited to highlight the semantic information of small target in the high-level features. Specially, top-down region-wise attention is designed to enrich the local semantics of lower-level features. Along these lines, lower-level features are incorporated with higher-level local information beyond the limitations of their receptive fields, but their spatial subtleties are preserved.

Suppose that lower-level features x_l contains C channels, and the size of feature map of each channel is $H \times W$. To decode the details of spatial features, the higher-level features

x_h is up-sampled to the same spatial resolution as x_l, and 1×1 convolution is further used to adjust the number of channels of x_h to C. The conversion process is as follows:

$$x_h^{'} = F_{c_1}(F_U(x_h)) \tag{8}$$

where $F_U(\cdot)$ is the up-sampling operation, $F_{c_1}(\cdot)$ is the 1×1 convolution.

Then, we sequentially generate fixed-size local regions centered on the pixels of $x_h^{'} \in R^{C \times H \times W}$, and calculate the respective average values of the different local regions, so that each descriptor can contain information of multiple dense local contexts. Specifically, the local patch feature Z_r^c of the c-th channel is calculated as follows:

$$Z_r^c = \frac{1}{s \times s} \sum_{i_s, j_s = 1}^{s \times s} Patch_{i,j}(x_h^{'}), (1 \le i \le W; 1 \le j \le H) \tag{9}$$

where the local semantic of each position (i, j) is derived from the average aggregation of the local patch generated by each position (i, j) in $x_h{}'$, and $s \times s$ represents the size of the local patch. In this way, a vector Z_r with C channels can be generated.

Inspired by SENet [35], the bottleneck gating is designed to learn the the attention vector $H(x) \in R^{C \times H \times W}$ of Z_r. Specially, it consists of two different convolution layers with different functions, and the gating mechanism for generating attention map is symbolized as follows:

$$H(x) = \sigma[BN(W_i \delta(BN(W_r Z_r)))] \tag{10}$$

where σ and δ denote sigmoid and ReLU functions, respectively. W_r and W_i represent the 1×1 convolution, W_r is used to reduce the feature dimension with the ratio r, and W_i is used to restore the feature dimension back to C. And BN is the batch normalization operation.

Then, the lower-level features $x_{h \to l} \in R^{C \times H \times W}$ of the local semantic modulation can be obtained via

$$x_{h \to l} = H(x) \otimes x_l \tag{11}$$

where $\otimes$ represents element-wise multiplication.

Meanwhile, the bottom-up point-wise attention is designed to enrich the semantic information of higher-level features with fine subtleties of lower-level features. In contrast to the top-down region-wise attention, this modulation pathway utilizes the point-wise channel interactions at each spatial location and propagates the local detail information in a bottom-up way. Specially, the modulation mechanism consists of two different PWConv [24] to aggregate channel feature context, and the attention vector $L(x) \in R^{C \times H \times W}$ of bottom-up modulation is calculated via a bottleneck gating as follows:

$$L(x) = \sigma(BN(\mathrm{PWConv}_2(\delta(BN(\mathrm{PWConv}_1(x_l)))))) \tag{12}$$

where σ means sigmoid function, δ means ReLU function. PWConv_1 and PWConv_2 have kernel sizes of $C/r \times C \times 1 \times 1$ and $C \times C/r \times 1 \times 1$, respectively. And BN is the batch normalization operation.

Then the higher-level features $x_{l \to h} \in R^{C \times H \times W}$ of local detail modulation can be obtained via

$$x_{l \to h} = L(x) \otimes x_h^{'} \tag{13}$$

where $\otimes$ represents element-wise multiplication.

In general, obtaining dense multi-scale information can enrich the feature representation of small targets. Therefore, it is necessary to explore different details and semantic information in different scale spaces. Different from the common multi-scale feature aggregation method [13–18] and the single-level asymmetric cross-layer feature aggregation method [24,25], we design a gated inverted pyramid to utilize multi-scale information in this work. Specially, as shown in Figure 5a, the top-down region-wise attention and bottom-up point-wise attention are applied to adjacent x_h and x_l to make the x_l enriched in

semantics and x_h is enriched in details, that is, semantic-guided detail features and detail-guided semantic features are generated simultaneously. Then both of them are combined to enhance the consistency between point-wise fine details and point-wise local semantics between adjacent scale features. Next, as shown in Figure 2b, $\{x_s\}_{s=1,\ldots,5}$ form the output of densely connected feature pyramid extraction module is taken as the first layer of the inverted pyramid, the adjacent feature maps of these five nodes are aggregated based on PAC in a pair-wise manner. Then, pair-wise modulated local context aggregation features of the first-level are generated. The calculation process is as following:

$$x_{45} = F_{c_3}((H(\rho(x_5^{'})) \otimes x_4) \oplus (L(x_4) \otimes x_5^{'})) \tag{14}$$

$$x_{34} = F_{c_3}((H(\rho(x_4^{'})) \otimes x_3) \oplus (L(x_3) \otimes x_4^{'})) \tag{15}$$

$$x_{23} = F_{c_3}((H(\rho(x_3^{'})) \otimes x_2) \oplus (L(x_2) \otimes x_3^{'})) \tag{16}$$

$$x_{12} = F_{c_3}((H(\rho(x_2^{'})) \otimes x_1) \oplus (L(x_1) \otimes x_2^{'})) \tag{17}$$

where $F_{c_3}(\cdot)$ is the 3×3 convolution, $\rho(\cdot)$ is the region-wise aggregation in Equation (9), $\otimes$ is element-wise multiplication, and $\oplus$ is element-wise summation.

We can regard x_{45}, x_{34}, x_{23} and x_{12} as each node in the second layer of the inverted pyramid generated by PAC, each of them includes two adjacent scales of information (i.e., one from the lower level features and the other from the higher level features). Subsequently, the x_{45}, x_{34}, x_{23} and x_{12} are further aggregated based on PAC, and pair-wise modulated local context aggregation features of the second-level are as follows:

$$x_{345} = F_{c_3}((H(\rho(x_{45}^{'})) \otimes x_{34}) \oplus (L(x_{34}) \otimes x_{45}^{'})) \tag{18}$$

$$x_{234} = F_{c_3}((H(\rho(x_{34}^{'})) \otimes x_{23}) \oplus (L(x_{23}) \otimes x_{34}^{'})) \tag{19}$$

$$x_{123} = F_{c_3}((H(\rho(x_{23}^{'})) \otimes x_{12}) \oplus (L(x_{12}) \otimes x_{23}^{'})) \tag{20}$$

where x_{345}, x_{234}, and x_{123} are regarded as three nodes in the third layer of the inverted pyramid, and each node contains three adjacent scales information.

Likewise, pair-wise modulated local context aggregation features of the third-level are generated, as follows:

$$x_{2345} = F_{c_3}((H(\rho(x_{345}^{'})) \otimes x_{234}) \oplus (L(x_{234}) \otimes x_{345}^{'})) \tag{21}$$

$$x_{1234} = F_{c_3}((H(\rho(x_{234}^{'})) \otimes x_{123}) \oplus (L(x_{123}) \otimes x_{234}^{'})) \tag{22}$$

where x_{2345} and x_{1234} are regarded as two nodes in the fourth layer of the inverted pyramid, and each node contains four adjacent scales information.

Then, pair-wise modulated local context aggregation features of the fourth-level is generated, as follows:

$$x_{12345} = F_{c_3}((H(\rho(x_{2345}^{'})) \otimes x_{1234}) \oplus (L(x_{1234}) \otimes x_{2345}^{'})). \tag{23}$$

Similarly, x_{12345} is regarded as one nodes in the fifth layer of the inverted pyramid, and each node contains five adjacent scales information.

(2) Hierarchical asymmetric combination

Similar to ACM [24], AFM [25] also modulates global semantics of high-level features and point-wise details of low-level features asymmetrically. The AFM merges cross-layer features in a hierarchical manner, but it neither fully explores feature matching in different spatial information fusion, nor designs local semantic signals of high-level features to modulate local details of cross-layer fusion features. In our work, to highlight the local presentation of small targets in cross-layer feature fusion, the convolution operation is first used to better match the compatibility of different spatial features, and then the local context

information of higher-level features and the point-wise details of lower-level features are designed to progressively modulate the fused features. As shown in Figure 5b, x_h is first transformed into $x_h^{'}$ through formula (8). Subsequently, $x_h^{'}$ and x_l are fused by element-wise addition and 3×3 convolution learning, and then $x_h^{'}$ is used to generate region-aware attention map to modulate cross-level fusion features. Next, x_l is applied to generate point-aware attention map to further modulate the cross-level fusion features embedded in the high-level regions, and finally the hierarchical asymmetric fusion features $x_{h \leftrightarrow l}$ is obtained. The conversion process of HAC is as follows:

$$x_{h \leftrightarrow l} = F_{c_3}((x_l \oplus x_h^{'})) \otimes H(x_h^{'}) \otimes L(x_l). \tag{24}$$

Subsequently, multi-scale features $\{x_s\}_{s=1,\ldots,5}$ are aggregated based on HAC in an inverted pyramid manner, as shown in Figure 2b. The inverted pyramid aggregation process based on HAC is similar to Equations (14)–(23) , and we will not describe it in detail here.

3.2.2. Recurrent Asymmetric Fusion

In PWAC, we have combined $\{x_s\}_{s=1,\ldots,5}$ to generate node features x_{45}, x_{34}, x_{23}, x_{12}, x_{345}, x_{234}, x_{123}, x_{2345}, x_{1234} and x_{12345} of different layers in the inverted pyramid. More importantly, x_1, x_{12}, x_{123}, x_{1234} and x_{12345} have the same size, and these features are located in different layers of the inverted pyramid, that is, they can be represented as $x_{12345} \rightarrow x_{1234} \rightarrow x_{123} \rightarrow x_{12} \rightarrow x_1$ from deep-level network to shallow-level network. Similarly, since different layer features have different subtle detail and semantic information, the feature associations among them are mined to better highlight the features of small infrared targets.

Inspired by the recurrent neural network [44], the RAF mechanism is designed, which recursively fuses multi-level feature in the inverted pyramid starts from x_{12345} to x_1, as shown in Figure 6. Obviously, the process of RAF is obviously different from the process of PWAC. Like the structure of PWAC, we also design two variants of recursive units in the RAF from different perspectives, namely RAF based on parallel gating and RAF based on hierarchical gating.

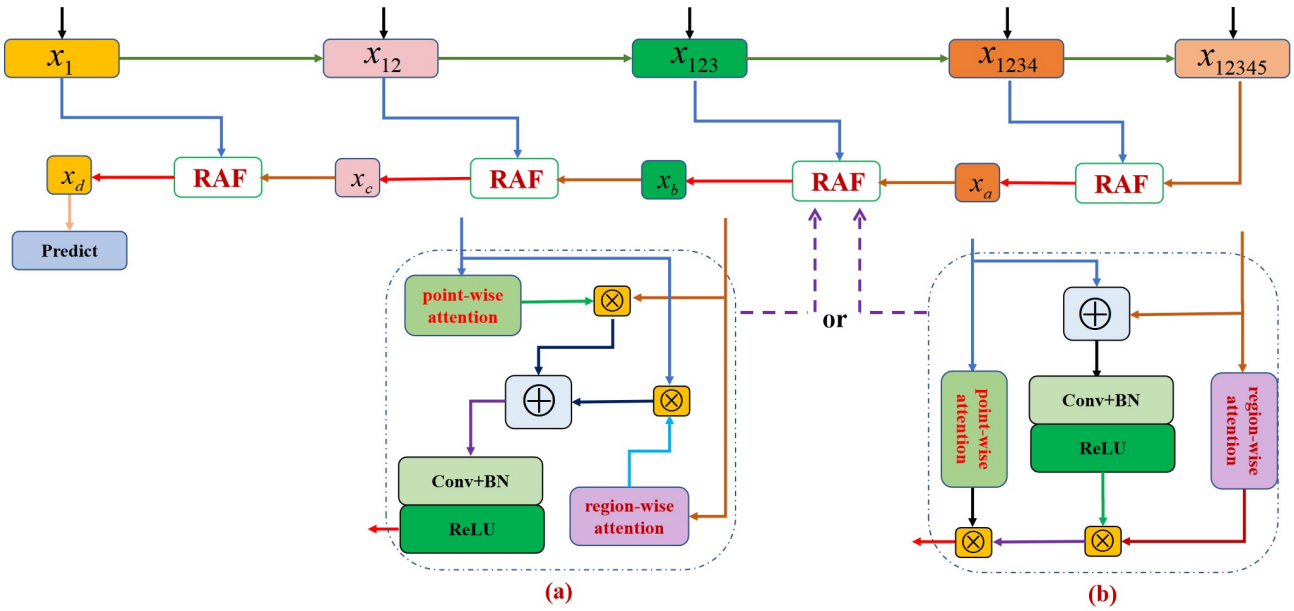


Figure 6. The illustration of the recurrent asymmetric fusion. (**a**) Recurrent asymmetric fusion based on parallel gating. (**b**) Recurrent asymmetric fusion based on hierarchical gating.

(1) RAF based on parallel gating

As shown in Figure 6a, the conversion process of RAF based on parallel gating is as follows:

$$x_a = F_{c_3}((H(\rho(x_{12345})) \otimes x_{1234}) \oplus (L(x_{1234}) \otimes x_{12345})) \tag{25}$$

$$x_b = F_{c_3}((H(\rho(x_a)) \otimes x_{123}) \oplus (L(x_{123}) \otimes x_a)) \tag{26}$$

$$x_c = F_{c_3}((H(\rho(x_b)) \otimes x_{12}) \oplus (L(x_{12}) \otimes x_b)) \tag{27}$$

$$x_d = F_{c_3}((H(\rho(x_c)) \otimes x_1) \oplus (\mathrm{L}(x_1) \otimes x_c)) \tag{28}$$

where $F_{c_3}(\cdot)$ is the 3×3 convolution, $\rho(\cdot)$ is the region-wise aggregation in Equation (9), $\otimes$ is element-wise multiplication, and $\oplus$ is element-wise summation.

(2) RAF based on hierarchical gating

As shown in Figure 6b, the conversion process of RAF based on hierarchical gating is as follows:

$$x_a = F_{c_3}((x_{12345} \oplus x_{1234})) \otimes H(\rho(x_{12345})) \otimes L(x_{1234}) \tag{29}$$

$$x_b = F_{c_3}((x_a \oplus x_{123})) \otimes H(\rho(x_a)) \otimes L(x_{123}) \tag{30}$$

$$x_c = F_{c_3}((x_b \oplus x_{12})) \otimes H(\rho(x_b)) \otimes L(x_{12}) \tag{31}$$

$$x_d = F_{c_3}((x_c \oplus x_1)) \otimes H(\rho(x_c)) \otimes L(x_1) \tag{32}$$

Subsequently, the convolution operation is employed to reduce the amount of channels to generate final spatial map x_d for the detection of dim and small infrared targets.

3.3. End-to-End Learning

In summary, an end-to-end APANet is proposed to explore the task of infrared dim and small target detection. Aiming at the serious category imbalance problem between background and small objects in infrared images, a Soft-IoU loss function [45] is adopted for this highly imbalanced object detection task. Given a sample image x, Θ represents the network parameters of the proposed APANet, which is defined as follows:

$$l_{soft\text{-}Iou}(x, s) = \frac{\sum_{i,j} p_{i,j} \cdot s_{i,j}}{\sum_{i,j} p_{i,j} + x_{i,j} - p_{i,j} \cdot s_{i,j}} \tag{33}$$

where $p = \sigma(APANet(x, \Theta)) \in R^{H \times W}$ denotes the final prediction map and $s \in R^{H \times W}$ denotes the labeled mask.

During network training, the parameter Θ is learned by minimizing the following total loss function over a given N training samples:

$$\Theta = \arg\min_{\Theta} \sum_{n=1}^{N} l_{soft\text{-}Iou}(\sigma(\mathrm{APANet}(x, \Theta)), s) \tag{34}$$

Obviously, APANet is an end-to-end optimized model with the aim of minimizing Equation (34).

4. Result

In this part, we first describe the benchmark dataset and evaluation strategy. Next, we describe the specific implementation details of the proposed method. Then, the proposed method is evaluated for quantitative and qualitative comparison with the currently most advanced infrared small target detection methods.

4.1. Dataset Description

The public SIRST dataset [24] is exploited to systematically evaluate the validity and robustness of the proposed APANet. This dataset consists of 427 images, all of which are

typical infrared small target detection images. Among them, the target occupies a small area of the entire image, and many targets are quite blurred and swamped in the complicate and messy background. More significantly, since the dataset does not contain images of successive frames, this makes it more difficult to improve the robustness of the model. For the SIRST dataset, we have randomly selected 341 images for model training and the remaining 86 images for model testing. In addition, the MDFA dataset [46] containing 10,000 training samples and 100 test samples is also used to evaluate the performance of the proposed APANet. All image samples in MDFA dataset are generated by the random combination of the real diversified background image and real small target or simulated them that obey Gaussian distribution. And the test samples do not contain any images in the training samples.

4.2. Evaluation Metrics

To more objectively and carefully evaluate the performance of the proposed APANet on two different datasets, the classic semantic segmentation evaluation metrics such as Precision, Recall, F-measure, and mean intersection over union (mIoU) are used [25]. F-measure can consider the Recall and Precision simultaneously, and it can be used as a reliable indicator to measure overall quality of segmentation. The F-measure is defined as:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{35}$$

As a pixel-level evaluation metric, mIoU can contour description capability of the model. The mIoU is defined as:

$$mIoU = \frac{A_t \cap A_d}{A_t \cup A_d} \tag{36}$$

where A_t and A_d represent the real target region and detected region, respectively.

In addition, the receiver operating characteristic (ROC) curve is exploited to express the dynamic relationship between true positive rate (TPR) and false positive rate (FPR). Meanwhile, the area under the curve (AUC) is also used as a key indicator to quantitatively evaluate ROC.

4.3. Implementation Details

In the proposed APANet, a backbone of pyramid feature extraction is constructed, which consists of 5 stages. Each stage consists of 3 convolutional layers, and all but the first stage are followed by a 2D average pooling layer to reduce the spatial resolution of features. The design details of different stages of backbone network in APANet are shown in Table 1.

Table 1. The design details of different stages of backbone in APANet.

Stage	Output	Backbone
Stage1	256×256	$[3 \times 3conv, 16] \times 3$
Stage2	128×128	$AvgPool2d$; $[3 \times 3conv, 32] \times 3$
Stage3	64×64	$AvgPool2d$; $[3 \times 3conv, 64] \times 3$
Stage4	32×32	$AvgPool2d$; $[3 \times 3conv, 128] \times 3$
Stage5	16×16	$AvgPool2d$; $[3 \times 3conv, 256] \times 3$

Our proposed APANet is evaluated on two public infrared small target detection datasets. In addition, the parameter of region size in top-down modulation of APANet is set to 8. In particular, the AdaGrad [47] optimizer is used to train the proposed APANet, and we set the initial learning rate to 0.05, the weight decay of 1×10^{-5}, and the batch size to 5. The size of the infrared image is resized to 256×256 pixels, and then they are input into the network. Moreover, the densely connected feature pyramid extraction module in our approach has two variants of DCMSF and rb-DCMSF, while the enhanced

asymmetric feature pyramid aggregation module including PWAC (i.e., PAC and HAC) and RAF (i.e., RAF based on parallel gating and RAF based on hierarchical gating) both contain two variants of parallel fusion and hierarchical fusion. In our method, PWAC and RAF simultaneously choose parallel gated fusion or hierarchical gated fusion, so our method has four variants, namely APANet-P (i.e., it consists of DCMSF, PAC, and RAF based on parallel gating), APANet-H (i.e., it consists of DCMSF, HAC, and RAF based on hierarchical gating), APANet-rb-P (i.e., it consists of rb-DCMSF, PAC, and RAF based on parallel gating), and APANet-rb-H (i.e., it consists of rb-DCMSF, HAC, RAF based on hierarchical gating), respectively.

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

To comprehensively verify the detection performance of APANet, quantitative evaluations and qualitative visualizations are performed on the benchmark dataset. Several mainstream methods in recent years are selected for comparison with APANet. First, we compare it with commonly used small infrared target detection methods based on model-driven design (i.e., IPI [9], MPCM [28], RLCM [29], FKRW [48], PSTNN [11], NRAM [49]). Table 2 presents the detailed hyperparameter settings for these model-driven methods. Moreover, we also compare it with recently proposed small infrared target detection methods based on data-driven CNN (i.e., ACM_FPN [24], ACM_U-Net [24], VGG16-FAMCA-LSPM [21], AGPCNet [25]).

Table 2. Hyper-parameters settings of the model-driven methods..

Methods	Hyper-parameter Settings
IPI [9]	Patch size: 50×50, sliding step: 10, $\varepsilon = 10^{-7}$, $\lambda = 1/\sqrt{\max(M, N)}$
MPCM [28]	Window size: 3×3, 5×5, 7×7, $K = 5$
RLCM [29]	$K_{th} = 5$, $K_1 = [2, 5, 9]$, $K_2 = [4, 9, 16]$
FKRW [48]	Window size: 11×11, $K = 4$, p = 6, $\beta = 200$,
PSTNN [11]	Patch size: 40×40, sliding step: 40, $\varepsilon = 10^{-7}$, $\lambda = 0.6/\sqrt{\max(n_1, n_2) * n_3}$
NRAM [49]	Patch size: 30×30, sliding step: 10, $\gamma = 0.002$, $\varepsilon = 10^{-7}$, $\lambda = 1/\sqrt{\min(M, N)}$

4.4.1. Quantitative Evaluation

Table 3 presents the experimental results for quantitative evaluation on SIRST dataset. As shown in Table 3, IPI outperforms our APANet-P on Recall and AUC metrics, but IPI is significantly inferior to our APANet-P on other evaluation metrics. Our APANet-P achieves the best effect on all other metrics except AUC and Recall. And on more representative mIoU and F-measure metrics, our APANet-P maintains the highest mIoU (0.7060) while achieving the highest F-measure (0.8277). The significant increase in these values indicates that our proposed APANet can mine discriminative features that are robust to diverse scenarios and can improve the accuracy of shape matching for detection of infrared small target. Moreover, the F-measure can evaluate the effect of the method more objectively because it comprehensively considers Precision and Recall. A single high precision or recall indicator cannot really achieve the desired results. For example, MPCM, RLCM, and IPI achieve the higher Recall, but these methods seriously sacrifice Precision. However, our proposed APANet can achieve a better balance between recall and precision. Overall, data-driven CNN methods achieve significant improvements over model-driven traditional methods. It is due to the fact that empirical setting of hyperparameters in traditional methods limits the generalization performance of these methods. While compared to data-driven CNN methods (i.e., ACM_FPN, ACM_U-Net, VGG16-FAMCA-LSPM, AGPCNet), APANet achieves significant improvements. This is attributed to our local-enhanced asymmetric pyramid aggregation module tailored for the detection of infrared small target. Among them, the detection effect of VGG16-FAMCA-LSPM is very poor, because the network parameters of this method are large, and more data is needed to train the network

better, which is obviously not suitable for training and evaluating the effect of small data sets. Of course, it also shows the importance of designing a reasonable network structure.

Table 3. Comparison with different detection methods on SIRST dataset.

Methods	Precision	Recall	mIoU	F-measure	AUC
MPCM [28]	0.1313	0.7181	0.1249	0.2220	0.8580
RLCM [29]	0.0406	0.7975	0.0402	0.0773	0.8933
FKRW [48]	0.0017	0.4688	0.0017	0.0034	0.6525
IPI [9]	0.2221	0.8959	0.2165	0.3560	0.9472
NRAM [49]	0.6452	0.5258	0.4079	0.5794	0.7628
PSTNN [11]	0.7431	0.6348	0.5205	0.6847	0.8173
ACM_FPN [24]	0.7098	0.6948	0.5411	0.7023	0.8761
ACM_UNet [24]	0.7359	0.7410	0.5854	0.7385	0.8880
VGG16-FAMCA-LSPM [21]	0.3392	0.4347	0.2354	0.3811	0.7171
AGPCNet [25]	0.8186	0.7546	0.6465	0.7853	0.8868
Ours					
APANet-P	0.8364	0.8192	0.7060	0.8277	0.9147
APANet-H	0.8089	0.8192	0.6864	0.8140	0.9133
APANet-rb-P	0.8143	0.8231	0.6930	0.8187	0.9185
APANet-rb-H	0.8141	0.7863	0.6666	0.7999	0.9071

In addition, we further evaluate the performance of APANet on the MDFA dataset, as shown in Table 4. As can be seen in Table 4, the data-driven CNN approach is also significantly better than the model-driven traditional approach. In addition, all four variants of our proposed APANet outperform the parallel asymmetric cross-layer fusion methods ACM_FPN and ACM_U-Net, and the multi-scale feature fusion method VGG16-FAMCA-LSPM. Compared with the hierarchical asymmetric cross-layer fusion method AGPCNet, the effect of APANet-P and APANet-H is inferior to that of AGPCNet on two more representative indicators, mIoU and F-measure. This is because the backbone network of AGPCNet adopts ResNet [50] architecture to extract features, while APANet-P and APANet-H only fuse features of different scales. Our APANet-rb-P and APANet-rb-H methods exploit the idea of residual design, and their performance is significantly better than that of AGPCNet on mIoU and F-measure. On the whole, our asymmetric local attention modulation method combining point-wise information and region-wise information is superior to the asymmetric attention modulation method based on global information and point-wise information.

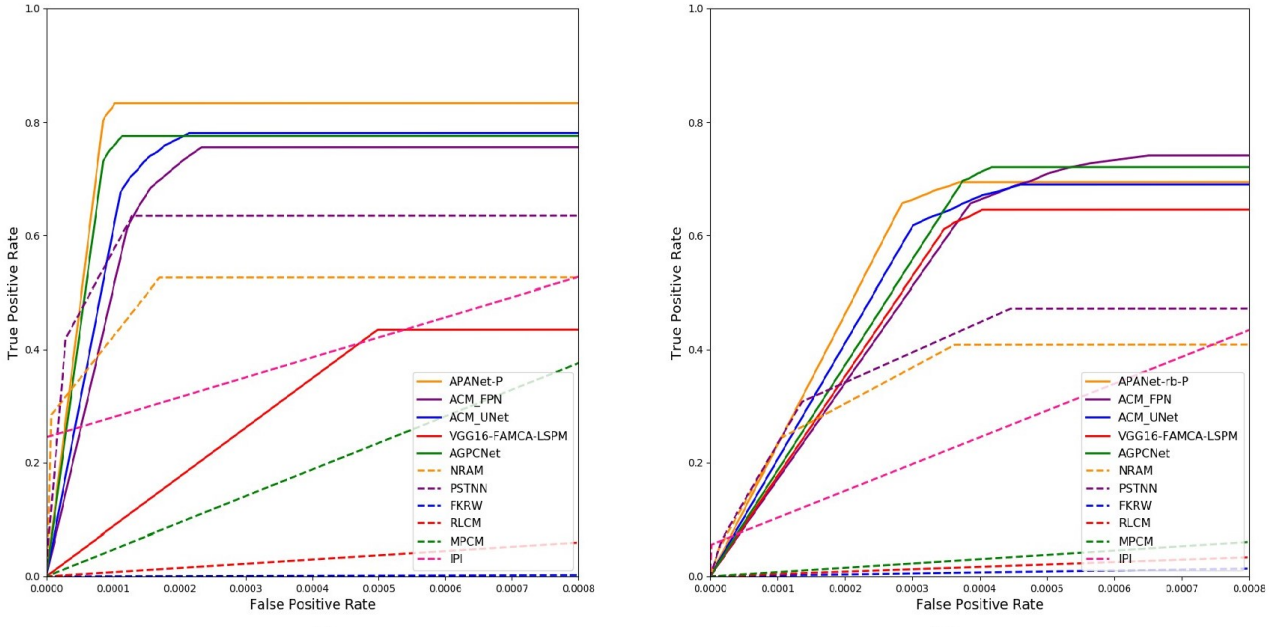
From Tables 3 and 4, it can be seen that the APANet-P and APANet-rb-P based on cross-layer parallel fusion perform better than the corresponding APANet-H and APANet-rb-H based on cross-layer hierarchical fusion. This suggests that in cross-layer feature fusion, parallel local asymmetric modulation can retain more intrinsic characteristics of the original features, and thus can better show the discriminant details of small infrared targets. On the SIRST dataset, the effects of APANet-P and APANet-H based on multi-scale feature fusion are better than those of the corresponding APANet-rb-P and APANet-rb-H based on multi-scale feature residual fusion. Nevertheless, on MDFA dataset, the results of APANet-rb-P and APANet-rb-H methods are better than those of APANet-P and APANet-H. Obviously, our densely connected multi-scale feature extraction method is more suitable for feature learning of small sample data. While our residual-based densely connected multi-scale feature extraction method is more suitable for feature learning of large-scale data with multiple kinds of backgrounds.

To further describe the effectiveness of our APANet, we also provide ROC curves obtained by different detection methods to visualize the comparison of AUC, as shown in Figure 7. Among them, we choose the best APANet method on the two indicators of

mIoU and F-measure as the benchmark for two different data sets. Obviously, on the SIRST dataset, APANet outperforms all comparative data-driven and model-driven infrared small target detection methods. On the MDFA dataset, the data-driven method is better than the model-driven method. However, the effect of APANet is roughly the same as that of two other data-driven methods, ACM_FPN and AGPCNet. This is because the MDFA dataset is a large-scale synthetic data with high noise, which seriously affects the discriminative feature learning of these data-driven deep networks. In addition, although the effect of APANet on AUC evaluation index is not obvious compared with ACM_FPN and AGPCNet, it still has more advantages in the evaluation of the two key indicators, mIoU and F-mesure. By comparing (a) and (b) in Figure 7, it can be seen that more real data sets and more accurate data annotation are more conducive to the network to learn discriminative features. In a word, a series of experimental results show that the proposed APANet has more advantages in background suppression, target detection and segmentation.

Table 4. Comparison with different detection methods on MDFA dataset.

Methods	Precision	Recall	mIoU	F-measure	AUC
MPCM [28]	0.0392	0.6439	0.0383	0.0738	0.8168
RLCM [29]	0.0318	0.6970	0.0313	0.0608	0.8403
FKRW [48]	0.0135	0.3851	0.0132	0.0260	0.6815
IPI [9]	0.2880	0.6290	0.2462	0.3951	0.8139
NRAM [49]	0.4669	0.4082	0.2784	0.4356	0.7039
PSTNN [11]	0.4520	0.4719	0.3002	0.4617	0.7358
ACM_FPN [24]	0.5247	0.7092	0.4318	0.6032	0.8748
ACM_UNet [24]	0.5780	0.6551	0.4431	0.6141	0.8440
VGG16-FAMCA-LSPM [21]	0.5673	0.6281	0.4246	0.5961	0.7757
AGPCNet [25]	0.5820	0.7098	0.4701	0.6396	0.8554
Ours					
APANet-P	0.5771	0.6800	0.4538	0.6243	0.8221
APANet-H	0.5507	0.7104	0.4498	0.6205	0.8579
APANet-rb-P	0.6162	0.6772	0.4763	0.6453	0.8469
APANet-rb-H	0.5862	0.7088	0.4724	0.6417	0.8589



(a) (b)

Figure 7. Illustration of ROC curve compared with other methods. (**a**) Comparison of different methods on SIRST dataset. (**b**) Comparison of different methods on MDFA dataset.

4.4.2. Qualitative Evaluation

The qualitative results obtained by different small target detection methods on some infrared example images are shown in Figure 8. Among them, green circles represent detection targets, while red circles represent false alarms. Meanwhile, Figure 9 shows the 3D visualization qualitative results of example images, ground truth, and different detection methods to facilitate the observation of target and clutter in the images. As shown in Figures 8 and 9, the model-driven traditional methods are prone to generate multiple false alarms and missing areas in complex scenes, because the traditional detection methods relies largely on the hand-crafted features extracted by artificial empirical design and cannot adapt the changes of the target size and scene categories. Compared with the traditional detection method, the CNN-based detection methods (i.e., ACM_FPN and AGPCNet) obtain better visualization results. However, ACM_FPN and AGPCNet also seem to generate some false alarms. Obviously, APANet-P is more robust for detecting small infrared targets in more scenarios, because it can locate more precise target position and segment more accurate target appearance. This is due to some different modules we designed can promote the APANet-P better to adapt the various changes of clutter background, target shape and target size, so that better detection and segmentation results can be obtained.

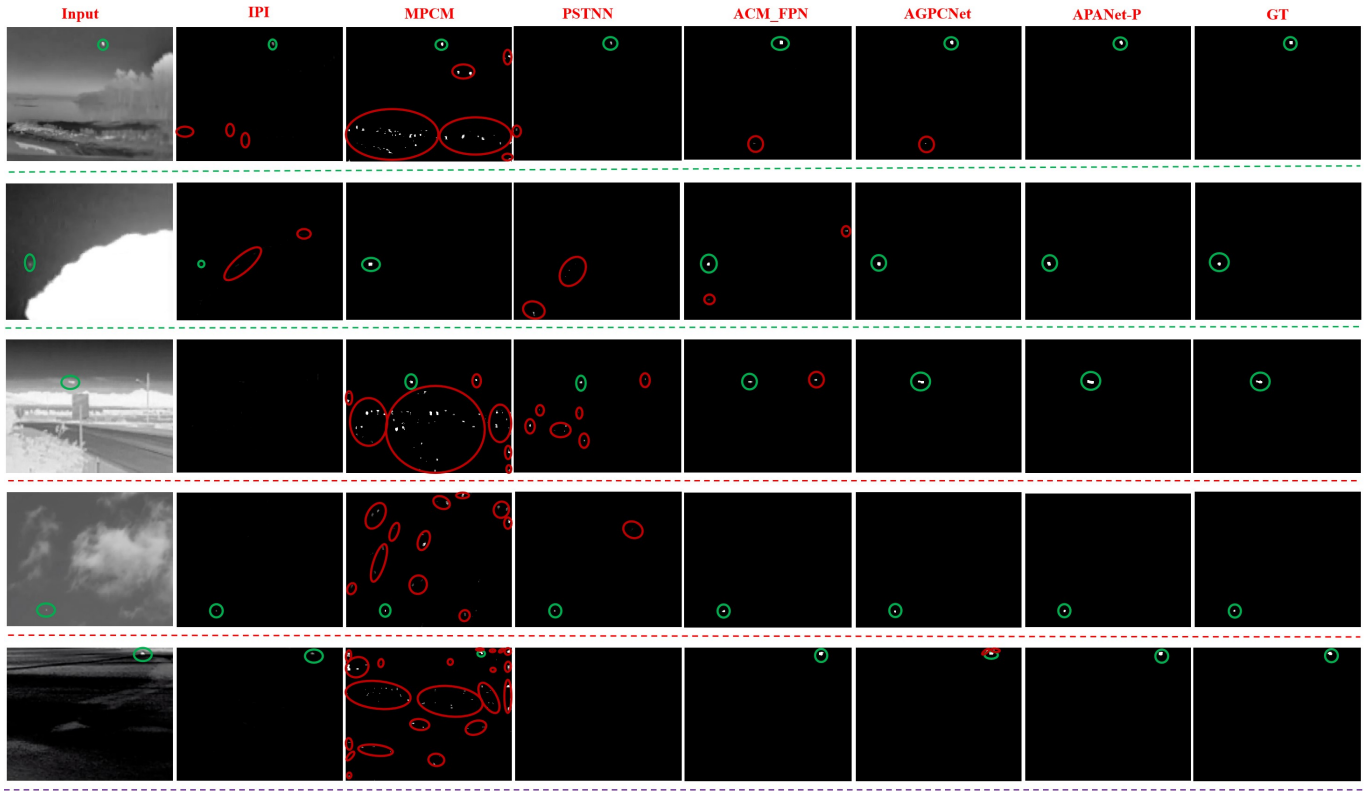


Figure 8. Qualitative outputs of different small infrared target detection methods.

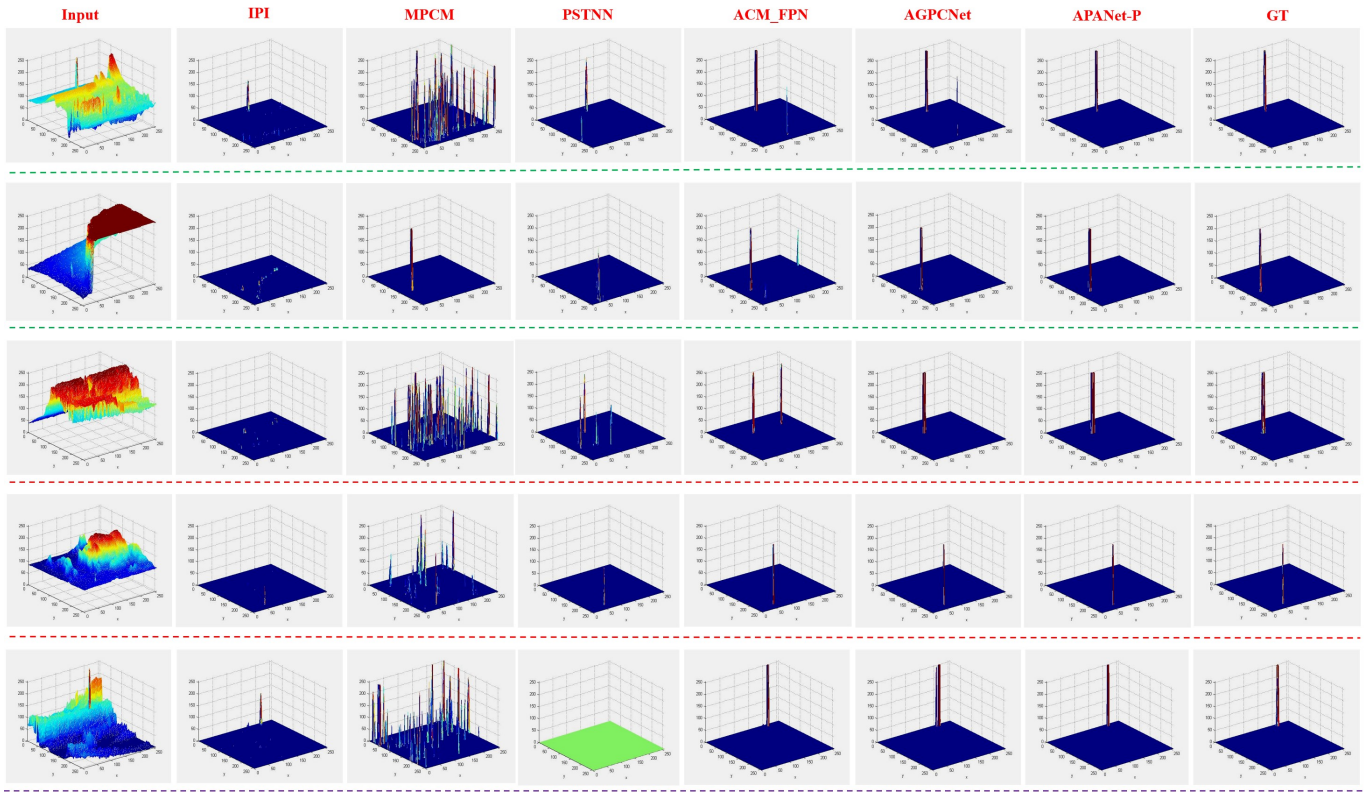


Figure 9. 3D visualization results of different small infrared target detection methods.

Furthermore, Figure 10 shows the visualization results obtained by four different variants of our APANet on some infrared example images. In order to present the segmentation outputs more intuitively and finely, we enlarge the target area to the lower right corner of the image. As shown in Tables 3 and 4, the overall effect of APANet-P is better than that of APANet-H. However, as shown in Figure 10, the design of APANet-H can detect dense multi-targets in more complex scenes, while the design of APANet-P will have some missed detections for dense multi-target scenes, which indicates that the design of hierarchical fusion is more suitable for dense multi-object scenes. In addition, our proposed APANet-P, APANet-H, APANet-rb-P and APANet-rb-H have different detection effects on different amounts of small infrared target datasets. As shown in Figure 10, the residual-based densely connected multi-scale feature extraction methods have more advantages than the densely connected multi-scale feature extraction methods in the accurate detection of the edge of the extended target with tens of pixels, while in the scene of multiple small target detection where the target is close to the surrounding background, it may bring some false alarms, or even split the whole target. Overall, although our proposed APANet can achieve good performance, it also has some limitations in some scene images that cannot segment the appearance contours of small infrared targets accurately.

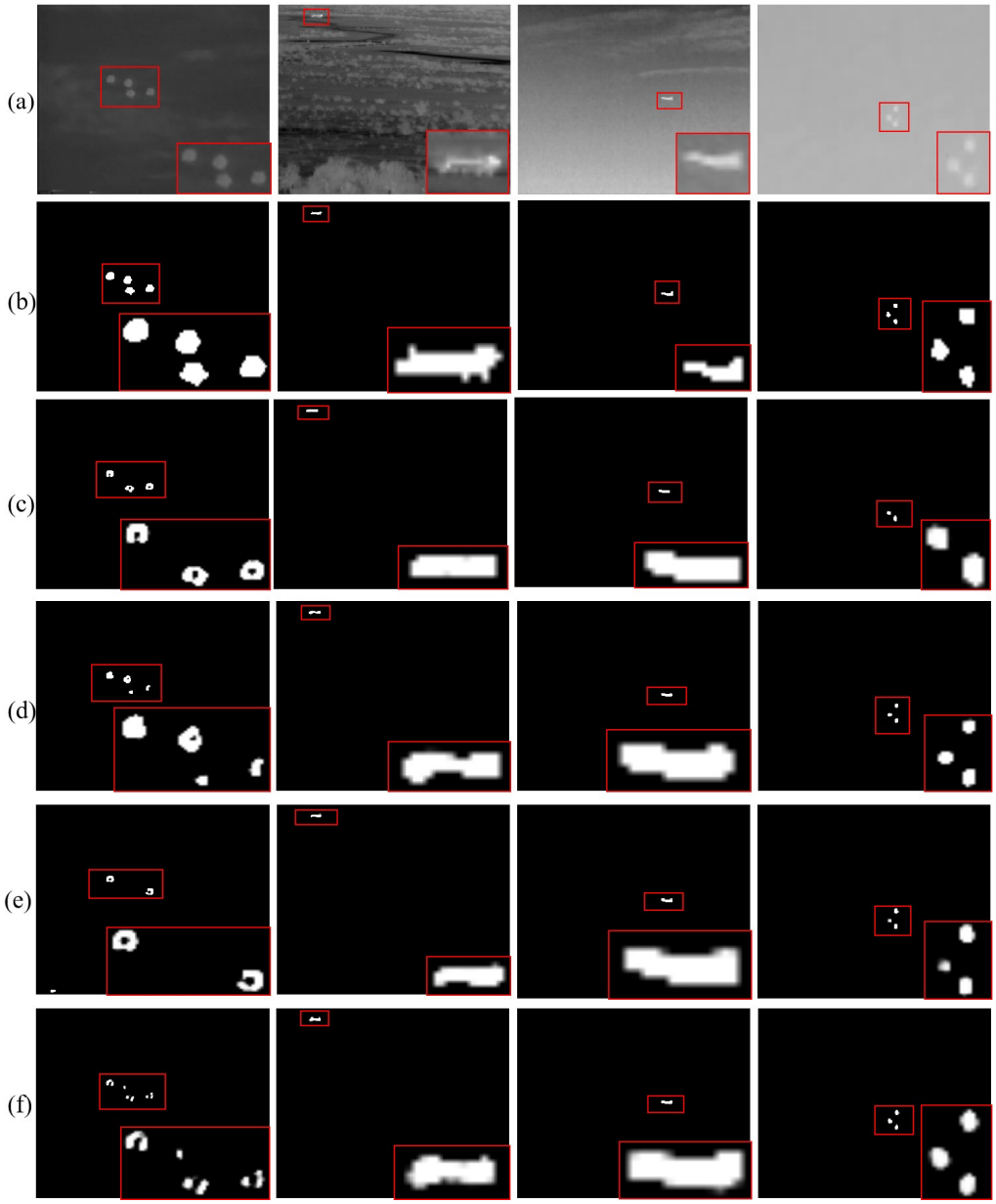


Figure 10. Segmentation results of proposed APANet on some infrared images. (**a**) The original image. (**b**) The ground truth. (**c**–**f**) The segmentation result of proposed APANet-P, APANet-H, APANet-rb-P, and APANet-rb-H, respectively.

5. Discussion

As mentioned earlier, our proposed APANet consists of some different components. In this section, we discuss our method in detail, and conduct ablation studies to demonstrate the effectiveness and contribution of each component to overall network model. Especially, using APANet-P as the baseline, and ablation analysis is conducted on the SIRST dataset. Firstly, the rationality of the densely connected CSDA design is verified, and then the effectiveness of different asymmetric feature learning is also verified. Furthermore, to conduct ablation experiments more fairly, we ensure that all parameter settings (e.g., image size, batch size, learning rate, optimizer, etc.) in ablation experiments are exactly the same.

All comparison methods of ablation analysis are as follows:

- APANet-P_without_DCCS: Remove the setting of dense connection based on CSDA of pyramid extraction module in APANet-P, that is, only the multi-scale features generated by the 5-stage convolutional layers are used for feature pyramid aggregation.
- APANet-P_without_RAF: Remove the setting of RAF of the asymmetric pyramid aggregation module in APANet-P, that is, only using the feature x_{12345} generated in the PWAC for small target detection.
- APANet-P_sum: Replace the setting of RAF of asymmetric pyramid aggregation module in APANet-P with the direct feature fusion, that is, the features summed by x_1, x_{12}, x_{123}, x_{1234} and x_{12345} are used for small target detection.
- APP-Net: Replace the settings of top-down region-wise attention and bottom-up point-wise attention of the asymmetric pyramid aggregation module in APANet-P with top-bottom point-wise attention and bottom-up point-wise attention.
- ALP-Net: Replace the settings of top-down region-wise attention and bottom-up point-wise attention of the asymmetric pyramid aggregation module in APANet-P with top-down region-wise attention and bottom-up region-wise attention.
- APANet-P-6: Modify the parameter size of the local region in the top-down region-wise attention of the asymmetric pyramid aggregation module in APANet-P to 6.
- APANet-P-10: Modify the parameter size of the local region in the top-down region-wise attention of the asymmetric pyramid aggregation module in APANet-P to 10.
- APANet-P-12: Modify the parameter size of the local region in the top-down region-wise attention of the asymmetric pyramid aggregation module in APANet-P to 12.
- APANet-P: Our proposed novel asymmetric pyramid aggregation network, which consists of DCMSF, PAC, and RAF based on parallel gating.

5.1. Effectiveness of Densely Connected Feature Extraction

In order to keep the spatial details of small targets in the features of the high-level network, a densely connected pyramid feature extraction mechanism combined with dual attention is designed in our APANet to transfer the attention-modulated shallow features to the attention-modulated deep features. Here, the design of dense cross-layer connections incorporating dual attention is removed to demonstrate that the design is useful for enriching feature representations of small targets. Compared with the APANet-P method, the APANet-P_without_DCCS method has a large performance degradation in the two indicators of mIoU and F-measure, as shown in Table 5. This shows that designing a densely connected dual attention mechanism in pyramid extraction module can effectively transfer the low-level large-scale spatial features in the network to the deep small-scale spatial features, so that the feature detail of dim and small targets can be better preserved in the multi-scale spaces at different layers.

5.2. Effectiveness of Pair-Wise Asymmetric Combination

In particular, pair-wise and asymmetric multi-layer feature combination based on top-down region-wise attention and bottom-up point-wise attention is also a key part of pyramid aggregation module in APANet. To demonstrate the effectiveness of asymmetric contextual modulation, two symmetric multi-layer feature fusion configurations are designed, namely APP-Net and ALP-Net. As shown in Table 5, the performance of APP-Net based on pixel-wise symmetric fusion and ALP-Net based on region-wise symmetric fusion shows performance degradation in the two indicators of mIoU and F-measure, while ALP-Net performs significantly better than APP-Net. These fully demonstrate that our designed pair-wise asymmetric multi-layer feature fusion is reasonable and effective. Our method is designed based on the interpretability research of deep networks, that is, low-level networks focus on the detail features, and high-level networks focus on the semantic features. Meanwhile, it also takes into account the particularity of infrared dim and small targets. Therefore, the feature modulation from higher-level network to lower-level network designs the region-wise perception mechanism, because local regions can contain more semantics than a single pixel, and region-wise perception is also applicable to the

local characteristics of infrared dim and small targets. However, low-level features adopt pixel-wise perception to focus on the detail information of target to modulate the high-level semantic features.

Table 5. Model ablation analysis on SIRST dataset.

Methods	mIoU	F-measure
Effectiveness of densely connected feature extraction		
APANet-P_without_DCCS	0.6444	0.7837
Effectiveness of pair-wise asymmetric combination		
APP-Net	0.6464	0.7852
ALP-Net	0.6829	0.8116
Effectiveness of recurrent asymmetric fusion		
APANet-P_without_RAF	0.6959	0.8207
APANet-P_sum	0.6865	0.8141
Influence of the size of the local region in region-wise attention		
APANet-P-6	0.6790	0.8088
APANet-P-10	0.6749	0.8059
APANet-P-12	0.6799	0.8095
Ours		
APANet-P	**0.7060**	**0.8277**

5.3. Effectiveness of Recurrent Asymmetric Fusion

The RAF mechanism is designed to fuse the multi-level spatial features of the same scale in the inverted pyramid generated by the PWAC mechanism to enrich the feature representation of small targets. Likewise, to demonstrate the effectiveness of RAF mechanism, two variants are designed, APANet-P_without_RAF and APANet-P_sum, respectively. APANet-P_without_RAF only utilizes the top-level feature generated by PWAC for small target detection, while APANet-P_sum simply fuses features from different layers generated by PWAC mechanism for detection of small target. As shown in Table 5, the methods of APANet-P_without_RAF and APANet-P_sum also show different degrees of performance degradation in mIoU and F-measure compared to the APANet-P method. This shows that recursively fusing the features generated by PWAC in the inverted pyramid can more obviously exploit the advantages of feature fusion at different layers. In addition, the APANet-P_sum method is even worse than the APANet-P_without_RAF method on the two indicators of mIoU and F-measure, which show that a reasonable feature fusion strategy can take advantage of multi-layer feature fusion. However, simple multi-layer feature fusion does not necessarily promote performance improvement, and even introduces noise to degrade performance.

5.4. Influence of the Size of the Local Region in Region-Wise Attention

The semantic features of high-level networks are richer and more discriminative. Given the unique characteristics of small objects, we try to embed the local semantics of higher-level features into lower-level features to enrich the semantics of low-level features. In order to evaluate the influence of the size setting of local regions in high-level features on small object detection, we conduct ablation experiments on the parameter of region size in top-down region-wise attention. The value of the region size in APANet-P is set to 8, and we separately change the value of the region size to 6, 10, and 12, which are called APANet-P-6, APANet-P-10, and APANet-P-12, respectively. Specially, it can be seen from Table 5 that the methods of APANet-P-6, APANet-P-10 and APANet-P-12 have different degrees of performance degradation in the two indicators of mIoU and F-measure compared to the APANet-P method. This shows that a reasonable setting of the parameter value of the region size in top-down region-wise attention can better detect the small infrared targets.

6. Conclusions

In this paper, different gated attention mechanisms are designed in our APANet framework to adaptively capture multi-scale information of target and effectively enhance local target features for small infrared target detection. Especially, APANet mainly contains two feature pyramid sub-modules with different roles. In the feature pyramid extraction module, multi-scale spatial features extractor containing densely connected dual attention mechanisms is designed to learn multi-scale detailed features of different layers of infrared targets, which can alleviate dim and small targets being lost in deeper networks. In the feature pyramid aggregation module, pair-wise asymmetric modulation with region-wise attention and point-wise attention is first constructed to merge cross-layer features of adjacent scales layer-wise until an inverted pyramid is generated. Moreover, the gated aggregation path of inverted pyramid can continuously emphasize the consistency between details and semantics of small target in multi-scale features, thus enhancing local response of small target and suppress complex background interference. Next, recursive asymmetric modulation is introduced to further improve the discrimination of small target along multi-level high-resolution features for final small infrared target detection. In addition, we have designed different variants for different components of APANet from different perspectives to better demonstrate the scalability of our methods.

Extensive experiments are conducted on two public datasets to illustrate that our APANet has the capacity to cope with small object detection tasks of complex scenes, and ablation studies also reveal the effectiveness of each module in our APANet. Therefore, the experimental results of our APANet can demonstrate the effectiveness of diverse cross-level local contextual modulation.

Author Contributions: G.L. designed the model, completed the experiments and wrote the paper. L.D. involved in model design and made important revisions. J.L. collected data and made some comments on the paper. W.X. provided technical guidance and partially financed the research. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported in part by the Fundamental Research Funds for the Central Universities of China under Grant 3132019340 and 3132019200. This paper was supported in part by high tech ship research project from ministry of industry and information technology of the peoples republic of China under Grant MC-201902-C01.

Data Availability Statement: The SIRST and MDFA dataset used for training and test are available at: https://github.com/YimianDai/sirstandhttps://github.com/wanghuanphd/MDvsFA_cGAN, accessed on 10 September 2022.

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

References

1. Prasad, D.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1993–2016. [CrossRef]
2. Rawat, S.; Verma, S.; Kumar, Y. Review on recent development in infrared small target detection algorithms. *Procedia Comput. Sci.* **2020**, *167*, 2496–2505. [CrossRef]
3. Zeng, M.; Li, J.; Peng, Z. The design of top-hat morphological filter and application to infrared target detection. *Infrared Phys. Technol.* **2006**, *48*, 67–76. [CrossRef]
4. 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]
5. Chen, C.; Li, H.; Wei, Y.; Xia, T.; Yan, Y. A local contrast method for small infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 574–581. [CrossRef]
6. 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.
7. 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]
8. Li, Y.; Li, Z.; Xu, B.; Dang, C.; Deng, J. Low-Contrast Infrared Target Detection Based on Multiscale Dual Morphological Reconstruction. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]

9. 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]
10. Wang, H.; Yang, F.; Zhang, C.; Ren, M. Infrared small target detection based on patch image model with local and global analysis. *Int. J. Image Graph.* **2018**, *18*, 1850002. [CrossRef]
11. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
12. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
13. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Cham, Switzerland, 2015; pp. 234–241.
14. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
15. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
16. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
17. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
18. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
19. He, J.; Deng, Z.; Zhou, L.; Wang, Y.; Qiao, Y. Adaptive pyramid context network for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 7519–7528.
20. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
21. Huang, L.; Dai, S.; Huang, T.; Huang, X.; Wang, H. Infrared small target segmentation with multiscale feature representation. *Infrared Phys. Technol.* **2021**, *116*, 103755. [CrossRef]
22. Wu, Z.; Su, L.; Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3907–3916.
23. Pang, Y.; Li, Y.; Shen, J.; Shao, L. Towards bridging semantic gap to improve semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4230–4239.
24. 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. 950–959.
25. Zhang, T.; Cao, S.; Pu, T.; Peng, Z. AGPCNet: Attention-Guided Pyramid Context Networks for Infrared Small Target Detection. *arXiv* **2021**, arXiv:2111.03580.
26. Deshpande, S.; Er, M.; Venkateswarlu, R.; Chan, P. Max-mean and max-median filters for detection of small targets. In Proceedings of the Signal and Data Processing of Small Targets 1999, Denver, CO, USA, 20–22 July 1999; International Society for Optics and Photonics: Bellingham, WA, USA, 1999; Volume 3809, pp. 74–83
27. Li, Y.; Li, Z.; Zhang, C.; Luo, Z.; Zhu, Y.; Ding, Z.; Qin, T. Infrared maritime dim small target detection based on spatiotemporal cues and directional morphological filtering. *Infrared Phys. Technol.* **2021**, *115*, 103657. [CrossRef]
28. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [CrossRef]
29. Han, J.; Liang, K.; Zhou, B.; Zhu, X.; Zhao, J.; Zhao, L. Infrared small target detection utilizing the multiscale relative local contrast measure. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 612–616. [CrossRef]
30. Liu, M.; Du, H.; Zhao, Y.; Dong, L.; Hui, M.; Wang, S. Image small target detection based on deep learning with SNR controlled sample generation. *Curr. Trends Comput. Sciene Mech. Autom.* **2017**, *1*, 211–220.
31. Shi, M.; Wang, H. Infrared dim and small target detection based on denoising autoencoder network. *Mob. Netw. Appl.* **2020**, *25*, 1469–1483. [CrossRef]
32. 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.
33. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
34. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
36. Woo, S.; Park, J.; Lee, J.; Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

37. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
38. Zhang, P.; Liu, W.; Wang, H.; Lei, Y.; Lu, H. Deep gated attention networks for large-scale street-level scene segmentation. *Pattern Recognit.* **2019**, *88*, 702–714. [CrossRef]
39. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
40. Li, G.; Yu, Y. Deep contrast learning for salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 478–487.
41. Zhang, P.; Wang, D.; Lu, H.; Wang, H.; Ruan, X. Amulet: Aggregating multi-level convolutional features for salient object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 202–211.
42. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.
43. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
44. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
45. Rahman, M.; Wang, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*; Springer: Cham, Switzerland, 2016; pp. 234–244.
46. 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, Seoul, Korea, 27 October–2 November 2019; pp. 8509–8518.
47. Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **2011**, *12*, 2121–2159.
48. 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]
49. 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]
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, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.