remote sensing

MDPI

Article

EAAU-Net: Enhanced Asymmetric Attention U-Net for Infrared Small Target Detection

Xiaozhong Tong, Bei Sun *, Junyu Wei, Zhen Zuo and Shaojing Su

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; tongxiaozhong@nudt.edu.cn (X.T.); yujy@nudt.edu.cn (J.W.); z.zuo@nudt.edu.cn (Z.Z.); ssjing@nudt.edu.cn (S.S.)

* Correspondence: sunbei08@nudt.edu.cn

Abstract: Detecting infrared small targets lacking texture and shape information in cluttered environments is extremely challenging. With the development of deep learning, convolutional neural network (CNN)-based methods have achieved promising results in generic object detection. However, existing CNN-based methods with pooling layers may lose the targets in the deep layers and, thus, cannot be directly applied for infrared small target detection. To overcome this problem, we propose an enhanced asymmetric attention (EAA) U-Net. Specifically, we present an efficient and powerful EAA module that uses both same-layer feature information exchange and cross-layer feature fusion to improve feature representation. In the proposed approach, spatial and channel information exchanges occur between the same layers to reinforce the primitive features of small targets, and a bottom-up global attention module focuses on cross-layer feature fusion to enable the dynamic weighted modulation of high-level features under the guidance of low-level features. The results of detailed ablation studies empirically validate the effectiveness of each component in the network architecture. Compared to state-of-the-art methods, the proposed method achieved superior performance, with an intersection-over-union (IoU) of 0.771, normalised IoU (nIoU) of 0.746, and F-area of 0.681 on the publicly available SIRST dataset.

Keywords: infrared small target detection; enhanced asymmetric attention mechanism; feature fusion; U-Net

check for updates

Citation: 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. https://doi.org/10.3390/rs13163200

Academic Editor: Angel D. Sappa

Received: 30 June 2021
Accepted: 10 August 2021
Published: 12 August 2021

1. Introduction

The detection of infrared small targets plays a critical role in infrared search and tracking systems, military early warning systems, remote sensing systems, and other applications owing to the ability of infrared radiation to penetrate obstacles such as fog and other atmospheric conditions and that of infrared sensors to capture images regardless of lighting conditions [1]. Such target detection systems employ passive sensors with low observability, and the all-weather, multi-scene nature of infrared imaging contributes to its suitability in a wide range of applications. With the rise in unmanned aerial vehicles, the use of infrared sensors for field detection and rescue has further enriched the application scenario of infrared target detection. However, because of the long infrared imaging distance and low-resolution, infrared targets usually have a low signal-to-noise ratio, and dim and small target profiles exhibit limited shape features and are, thus, easily submerged in strong noise and background cluster, as shown in Figure 1. Because of these attributes, infrared small target detection remains challenging and has attracted considerable attention.

Infrared small target detection methods can be broadly divided into two categories: single-frame detection and multi-frame detection [2]. Single-frame detection uses target and background information from a single image only, whereas multi-frame detection is based on dynamically changing information contained in multiple images. The latter tends to be inefficient and is often unable to perform real-time end-to-end detection [3]. Traditional methods for detecting infrared small targets based on single-frame detection include filter-based methods [4,5], local contrast-based methods [6–9], and low-rank-based methods [10–13]. However, these traditional methods based on handcraft fixed sliding windows, step sizes, and fixed hyperparameters are incapable of detecting targets accurately when the characteristics of the real scene (e.g., target size, shape, and background clutter) change significantly from those expected.

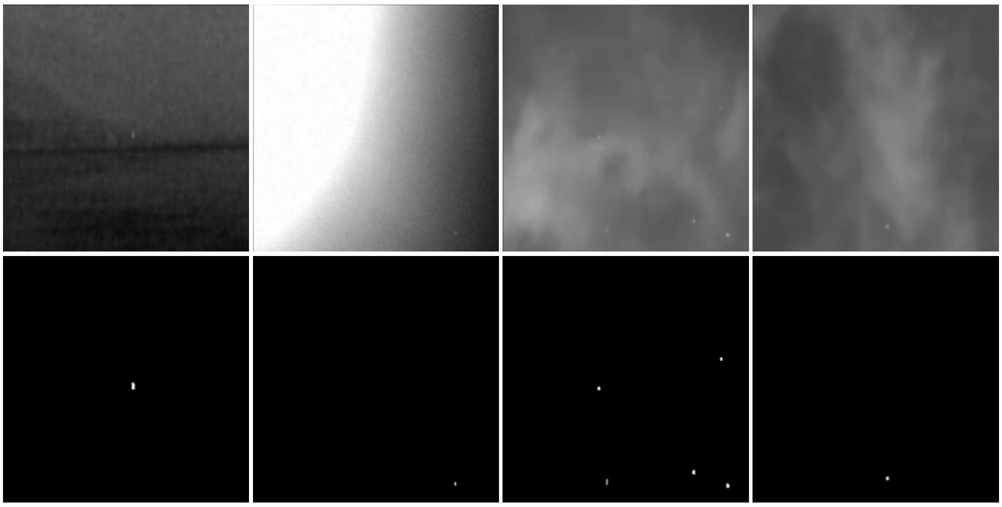


Figure 1. Examples of infrared small targets submerged in complex backgrounds that are difficult to detect accurately.

In recent years, with the success of deep learning in the field of computer vision, convolutional neural networks (CNNs) have been applied in infrared small target detection. In contrast to traditional methods, CNN-based methods learn the features of infrared small targets through the training of neural networks with large amounts of data. Liu et al. [14] were the first to propose the use of CNNs for infrared small target detection. Gao et al. [15] subsequently proposed a high-precision dim and small target detection algorithm based on feature mapping with a spindle network structure. McIntosh et al. [16] proposed a network that optimises a ‘target to clutter ratio’ (TCR) metric defined as the ratio of the output energies produced by the network in response to targets and clutter and compared it to state-of-the-art detectors such as Faster-RCNN [17] and Yolo-v3 [18]. Dai et al. [19] presented a typical infrared small target dataset, SIRST, and designed an asymmetric contextual module (ACM) to obtain richer semantic information and encode spatial details in infrared small target detection. Hou et al. [20] combined handcrafted feature methods with a CNN feature extraction framework, established a mapping network feature maps and the likelihood of small targets in the image, and applied thresholds on the likelihood maps to segment real targets. Fang et al. [21] integrated global and local dilated residual convolution blocks into U-Net for the remote infrared detection of unmanned aerial vehicles (UAVs). Huang et al. [22] used multiple well-designed local similarity pyramid modules (LSPMs) and attention mechanisms for the segmentation of infrared small targets. Although recent CNN-based approaches have achieved some performance improvements, they still suffer from target loss and unclear segmentation of details. Recent research has shown that a ’channel shuffle‘ fusion of spatial attention and channel attention can highlight the optimal semantic feature regions and reduce the computational complexity [23], thereby helping to preserve and extract the primitive features of small targets in deep networks.

Based on the above analysis, for features such as class imbalance and weak texture between infrared small targets and an image background, target detection networks need to focus not only on the spatial and channel features within a given image layer, but also on the feature connections between different layers to achieve cross-layer feature fusion. Inspired by channel shuffle units [23] and cross-layer feature fusion [24], we designed an enhanced asymmetric attention (EAA) module to improve performance without increasing the required network parameters or the model complexity, using cross-layer feature fusion and spatial and channel information exchange between the same layers to focus the network on the detailed content and spatial location information of small targets.

The main contributions of this paper are as follows.

(1) We propose EAAU-Net, a lightweight network for single-frame infrared small target detection, and experimentally demonstrate its ability to effectively segment the details of images of small targets and obtain satisfactory results.

(2) We present an EAA module designed to not only focus on spatial and channel information within layers, but also to apply cross-layer attention from shallow to deep layers to perform feature fusion. This module dynamically senses the fine details of infrared small targets and processes detailed target information.

(3) Experiments on the SIRST dataset show that our proposed EAAU-Net has the capacity to achieve better performance than existing methods and exhibits greater robustness to complex background clutter and weak texture information.

The remainder of this paper is organised as follows. Section 2 provides a brief review of related work. Section 3 describes the architecture of EAAU-Net in detail. Section 4 describes the experiments conducted and analyses the results obtained. Section 5 presents concluding remarks.

2. Related Work

In this section, we briefly review infrared small target detection and the existing methods relevant to this work.

2.1. Infrared Small Target Detection

Infrared small target detection has been investigated for decades. Traditional infrared small target detection methods measure the discontinuity between a target and its background. Such methods based on single-frame detection include two-dimensional least mean square (TDLMS) adaptive filtering, top-hat filtering, and maximum mean filtering [1]. However, these methods are easily affected by background clutters and noise [25]. In addition, human vision system (HVS)-based methods [6,7,26,27] have also been introduced to perform infrared small target detection owing to their ability to quickly extract valid information from complex scenes [28]. However, HVS-based methods are also susceptible to factors such as background noise, which affects the detection performance.

In recent years, CNN-based methods for infrared small target detection have attracted considerable attention. CNN-based methods with powerful model-fitting capabilities have achieved better performance than traditional methods by learning small target features in a data-driven manner. Deng et al. proposed a multi-scale CNN [29] for feature learning and classification. They used a network called MCNN to automatically extract the features of objects at multiple time scales and frequencies and combined low-level features with high-level features for spatial infrared point target recognition. Shi et al. proposed an end-to-end infrared small target detection model called CDAE based on denoising autoencoder networks and CNNs [30]. They treated small targets in infrared images as 'noise' and transformed the small target detection problem into a denoising problem to achieve better detection. Inspired by the U-Net [31] segmentation network, Zhao et al. [25] proposed a CNN for semantic segmentation of images containing infrared small targets by combining semantic constraint modules and implemented real-time infrared small target detection in an actual field scene. Zhao et al. proposed a generative adversarial network (GAN)-based detection model [32] to detect the basic characteristics of infrared small targets. In addition, a series of deep learning networks [15,16,33,34] have been designed to recognise the basic features of infrared small targets, with the aim of improving the detection accuracy. The attention mechanism has been demonstrated to enhance the contextual information of features by focusing the network on key areas [35]. Dai et al. [24] used bottom-up pixel-by-pixel convolution across layers to dynamically perform weight modulation of high-level features using low-level information embedded in high-level coarse feature maps and demonstrated in [19] that asymmetric modulation modules yield richer semantic information and spatial detail encoding. However, despite recent improvements in network performance, significant target loss remains at the deep level. Furthermore, these algorithms are less robust against scenarios such as complex backgrounds, low signal-to-noise ratios, and dim targets.

2.2. Attention and Cross-Layer Feature Fusion

Deeper networks can extract richer semantic information regarding targets and improve their learning capability over shallower networks, but small targets are easily lost in deeper networks. Consequently, the detection of small targets has been a key challenge in general-purpose computer vision. In particular, for infrared small target images, aspects such as low target signal-to-noise ratios, absence of texture and shape features, and the fact that the pixels of an infrared small target may occupy only 0.1% or less of a sensor image area further increase the challenge of extracting small target features in deep networks. To increase performance, anchor matching strategies have been elaborated [36], networks have been trained using scalable schemes [37], and coding contexts have also been extensively explored; this provides evidence beyond the object through the extraction and connection of features in a magnified window around the object [38], thereby alleviating the problems caused by small objects [39,40].

In addition, reawakening strategies involving the recall of prior or low-level feature information have been investigated [41]. Luong et al. [42] proposed the use of global and local attention for neural machine translation. Global attention refers to all prior memories, i.e., all previous feature maps, whereas local attention refers to several prior memories in a current prediction, i.e., several previous feature maps in the current layer. The perceptual area of the network is usually expanded through the use of multi-scale attention mechanisms designed for this purpose.

For accurate object localisation and segmentation, U-Net [31] and feature pyramid networks (FPNs) [43] are the classical networks for semantic segmentation. They follow a rough-to-fine strategy to hierarchically combine subtle features from lower layers and coarse semantic features from higher layers. However, most studies focus on the construction of complex paths to span the features at each layer [44]. Feature fusion methods via summation or cascading do not provide the network with the ability to dynamically select relevant features from lower layers. The bottleneck in infrared small target detection involves the retention and highlighting of the features of dim and small targets in deeper layers. Recently, methods have been proposed [45,46] for modulating low-level features in skip connections through a global channel attention module [47] using high-level features as a guide. Dai et al. [19] proposed ACM modules following the idea of cross-layer modulation, using bidirectional paths (top-down and bottom-up) to retain prior feature information for cross-layer feature fusion. They also demonstrated that bottom-up cross-layer feature fusion with pointwise attention modulation in CNNs can preserve and highlight the fine details of infrared small targets. Zhang et al. [23] achieved feature fusion and improved the target detection performance using spatial and channel attention mechanisms by implementing spatial and channel information exchange of target features through a shuffle unit.

3. Proposed Method

In this section, we describe EAAU-Net in detail. As shown in Figure 2, the network comprises three components: encoder, EAA module, and decoder. The following subsections detail the principal building blocks, overall architecture, and training loss function of the proposed EAAU-Net.

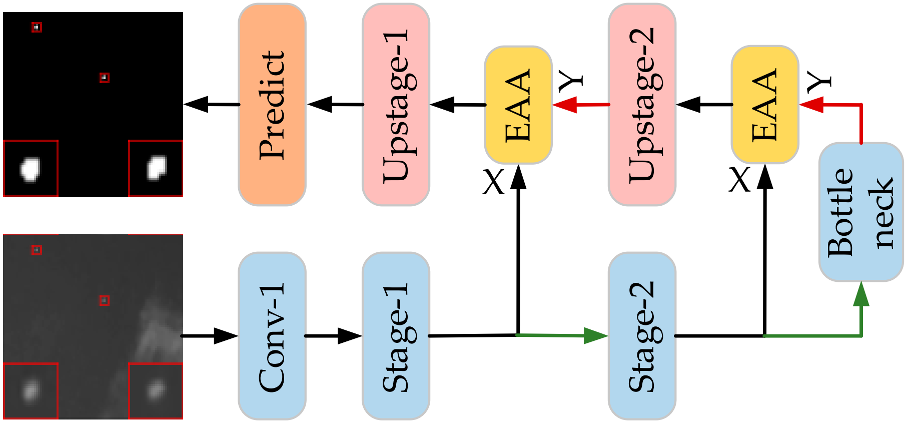


Figure 2. The proposed EAAU-Net, which incorporates the lower encoding module, the upper decoding module, and the enhanced asymmetric attention (EAA) module into a U-Net. The green line and red line represent the down-sampling and up-sampling operator, respectively.

3.1. Network Architecture

As shown in Figure 2, EAAU-Net takes a SIRST image as input, performs down-sampling, asymmetric attention feature fusion, and up-sampling operations in succession, and finally, outputs the segmentation result through the prediction module. The proposed architecture mainly consists of two components—U-Net [31] as a host network, with the proposed EAA module performing the cross-layer feature fusion operation; ResNet-20 [48] as the backbone architecture—as shown in Table 1. The deeper layers of the network are designed to be able to extract richer semantic information, as spatially finer shallow features and semantically stronger deeper features are considered crucial for detecting infrared small targets [34]. U-Net has natural advantages for infrared small target detection. Among these advantages, shallow network information ensures the integrity and extractability of the target information feature, and the lightweight network structure improves the inference efficiency. In addition, it is possible to pass high-resolution information throughout the network by skipping connections. This method has been used in detection networks [25,32].

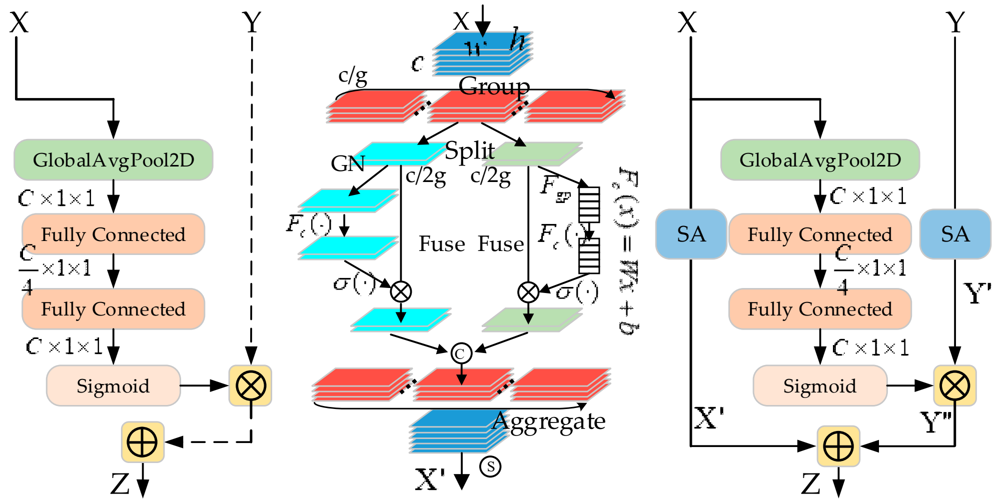
Table 1. EAAU-Net backbones.

Stage	Output	Backbone
Conv-1	480 × 480	$3 \times 3\text{conv}, 16$
Stage-1/UpStage-1	480 × 480	$\begin{bmatrix} 3 \times 3\text{conv}, 16 \\ 3 \times 3\text{conv}, 16 \end{bmatrix} \times \text{b}$
Stage-2/UpStage-2	240 × 240	$\begin{bmatrix} 3 \times 3\text{conv}, 32 \\ 3 \times 3\text{conv}, 32 \end{bmatrix} \times \text{b}$
Bottleneck	120 × 120	$\begin{bmatrix} 3 \times 3\text{conv}, 64 \\ 3 \times 3\text{conv}, 64 \end{bmatrix} \times \text{b}$

ResNet-20 is used as the backbone architecture to enhance the learning of the CNN, further improving the network's ability to fully exploit different levels of features in down-sampling and up-sampling phases, preserving fine features as well as deep semantic information, and enhancing spatial information and feature propagation. In addition, to study the relationship between the performance and network depth, we set ResNets of different depths (block number b in each stage) in experiments conducted. For b = 3, the network used the standard ResNet-20 backbone [48]. In Table 1, only the first convolutional layer of Stage-2 and Stage-3 is sub-sampled.

3.2. *Enhanced Asymmetric Attention (EAA) Module*

The up-sampling (decoding) process fuses the feature map information output using the encoder module by skipping connections, as well as using additional contextual and spatial information of the feature map from the low-resolution decoder block. We propose an EAA module, which mainly consists of two components—a bottom-up asymmetric attention (BAA) block and a shuffle attention (SA) block, as shown in Figure 3. The BAA block enables cross-layer feature fusion and highlights the fine details of a target. The SA block focuses on spatial and channel feature information within layers through channel shuffling, captures the spatial and channel correlation between features based on contextual information and weight selection regions, and enhances the spatial and channel information exchange of features to improve the performance of the proposed approach. In this subsection, we present the details of the entire module.



(a) BAA block (b) SA block (c) EAA

Figure 3. The proposed modules. (**a**) Cross-layer bottom-up asymmetric attention (BAA) block. (**b**) Shuffle attention (SA) block. (**c**) Enhanced asymmetric attention (EAA) module. ⊗, ©, Ⓢ and σ(·) denote element-wise product, concat, channel shuffle, and sigmoid, respectively.

3.2.1. Bottom-Up Asymmetric Attention (BAA) Block

Deep networks can provide high-level semantic features and understanding of scene context, but inevitably increase the risk of losing the spatial details of the target [44]. Inspired by bottom-up local attention modulation techniques [24], we aim to aggregate global contextual information by adding a global averaging pool to dynamically perceive the fine details of infrared small targets through low-level features. Therefore, we use a global channel attention approach BAA block (as shown in Figure 3a), in contrast to the traditional top-down path, to embed smaller-scale details into high-level coarse feature maps, with dynamic weighted modulation of the high-level features guided by low-level features. By default, X and Y refer to the low-level fine detail information and deep semantic information, respectively. For a given feature map output, both have the same size after the up-sampling and convolution operations. To preserve and emphasise the detailed information of the infrared small targets, the global channel attention module G first employs global averaging pooling in the bottom-up modulation path to aggregate the global contextual information, resulting in the following channel statistics:

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

where H and W denote the height and width of each feature map, respectively. The bottom-up modulation weight $\mathrm{G}(\mathrm{X}) \in \mathbb{R}^{C}$ can be expressed as

$$\mathrm{G}(\mathrm{X}) = \sigma(\mathcal{B}(\mathrm{W}_2\delta(\mathcal{B}(\mathrm{W}_1 x)))) \tag{2}$$

where C, σ, $\mathcal{B}$, and δ denote the number of channels, sigmoid function, batch normalisation (BN), and rectified linear unit (ReLU), respectively. W_1 denotes the first fully connected layer, which aims to reduce the number of channels, and W_2 denotes the second fully connected layer, which aims to recover the number of channels. For the relatively high-level feature map Y, the bottom-up modulation of the feature map by the global channel attention module can be expressed as

$$\mathrm{Z} = \mathrm{G}(\mathrm{X}) \otimes \mathrm{Y} \tag{3}$$

where $\otimes$ denotes the element-wise multiplication, and $\mathrm{G}(\cdot)$ denotes the bottom-up global channel attention modulation module.

3.2.2. Shuffle Attention (SA) Block

Attention mechanisms serve to improve the representation of interest, i.e., the ability of a model to focus on essential features while suppressing unnecessary features [35]. Zhang et al. [23] proposed a shuffle unit that efficiently combines spatial and channel characteristics to reduce the number of network parameters and improve performance. We incorporate this module in the encoding and decoding path of the proposed U-Net-based framework (as shown in Figure 3b), aiming to fully exploit the feature information and its correlation in the spatial and channel dimensions of both the low- and high-level networks to suppress possible noise while highlighting the optimum target feature regions.

For a given input feature map $X \in \mathbb{R}^{C\times H\times W}$, C, H, and W denote the number of channels and the height and width of each feature map, respectively. First, SA divides X into G groups along the channel dimension, i.e., $X = [X_1, \cdots, X_G], X_k \in \mathbb{R}^{C/G\times H\times W}$, where each sub-feature X_k progressively captures a specific semantic response during the training process, generating a corresponding importance factor for each sub-feature through the attention module. The input of X_k is split into two branches along the channel dimension, i.e., $X_{k1}, X_{k2} \in \mathbb{R}^{C/2G\times H\times W}$. The input grouped feature map X_{k1} first generates the channel statistics $s \in \mathbb{R}^{C/2G\times 1\times 1}$ by simply embedding the global information using global averaging pooling (GAP) and, then, multiplies s by the input grouping feature map X_{k1} pixel by pixel after processing with the enhancement function $\mathcal{F}_c \in \mathbb{R}^{C/2G\times 1\times 1}$ and sigmoid function to obtain the channel attention feature map $A_{\mathrm{c}} \in \mathbb{R}^{C/G\times H\times W}$. The channel attention is calculated via Equations (4) and (5):

$$s = \mathcal{F}_{gp}(X_{k1}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{k1}(i,j) \tag{4}$$

$$A_c = X'_{k1} = \sigma(\mathcal{F}_c(s)) \cdot X_{k1} = \sigma(W_1 s + b_1) \cdot X_{k1} \tag{5}$$

where $W_1, b_1 \in \mathbb{R}^{C/2G\times 1\times 1}$ are the parameters used to scale and shift s, and σ denotes the sigmoid function.

The input grouped feature map X_{k2} is first processed by the group norm (GN) [49], then, by the enhancement function $\mathcal{F}_c \in \mathbb{R}^{C/2G\times 1\times 1}$, after which it is multiplied pixel by pixel by the input grouped feature map X_{k2} to obtain the spatial attention map $A_{\mathrm{s}} \in \mathbb{R}^{C/G\times H\times W}$. The spatial attention calculation process can be expressed using Equation (6):

$$A_{\mathrm{s}} = X'_{k2} = \mathcal{F}_c(\mathrm{GN}(X_{k2})) \otimes X_{k2} = \sigma(W_2 \cdot GN(X_{k2}) + b_2) \cdot X_{k2} \tag{6}$$

where W_2 and b_2 are parameters with shape $\mathbb{R}^{C/2G\times 1\times 1}$, and σ denotes the sigmoid function.

After the input feature map has passed through the channel attention module and the spatial attention module, the network is able to focus on the feature map in which 'what' and 'where' are meaningful. The two branches from the channel features and the spatial features are connected such that the number of channels is equal to the number of inputs.

$$X'_k = [X'_{k1}, X'_{k2}] \in \mathbb{R}^{C/G \times H \times W} \tag{7}$$

Finally, all the sub-features are aggregated and a 'channel shuffle' operator similar to ShuffleNetv2 [50] is used to achieve cross-group information flow along the channel dimension, with the final SA module output being the same size as the input feature map.

3.2.3. EAA Module

To address the increased risk of losing the details of the target as the network deepens, and also the problem of insufficient exchange of feature information within layers, we propose an EAA module. In contrast to [19,24], we wish to fully exploit the feature information of both lower and higher layers, as is possible in a limited number of network layers, and ensure the spatial and channel information exchange in the same layer network. In particular, we hope to introduce BAA blocks in the same-layer information exchange to encode smaller-scale visual details to a deeper level and enable the exchange of high-level semantic and low-level detail feature information so as to ensure that the fine details of infrared small targets are not drowned out by background noise in the high-level feature information exchange. Finally, the multi-scale feature maps from different layers are intelligently fused to recover the full spatial resolution of the network output by iteratively fusing the information exchange in the lower spatial and channel dimensions as well as the information exchange feature maps in the higher spatial and channel dimensions modulated by the local attention across layers.

Given a low-level feature X and a high-level feature Y, the input feature map is first processed through the SA module to obtain X*'* and Y*'*. Then, the relatively high-level feature map Y, which is modulated by the bottom-up global channel attention module G, can be obtained using Equation (8):

$$\mathrm{Y}'' = \mathrm{G}(\mathrm{X}) \otimes \mathrm{Y}' \tag{8}$$

where $\otimes$ denotes element-wise multiplication, G(X) is reconstructed and broadcast to size $C \times H \times W$ and, then, pixel-by-pixel dot product, and finally, the output feature map is added to obtain the cross-layer fusion feature $Z \in \mathbb{R}^{C \times H \times W}$.

$$\mathrm{Z} = \mathrm{X}' + \mathrm{Y}'' = \mathrm{SA}(\mathrm{X}) + \mathrm{G}(\mathrm{X}) \otimes \mathrm{SA}(\mathrm{Y}) \tag{9}$$

where $\otimes$, $\mathrm{SA}(\cdot)$, and $\mathrm{G}(\cdot)$ denote element-wise multiplication, the shuffle attention block operation, and the bottom-up asymmetric attention block operation, respectively.

The EAA module enables the intelligent fusion of multi-scale feature maps at different stages, restoring the full spatial resolution of the network output by iteratively fusing coarse high-level semantic feature maps and fine low-level detail feature maps. The specific structure of the asymmetric attention feature fusion module is shown in Figure 3c; feature maps X and Y are processed by the shuffle unit, which fully exploits the spatial and channel information exchange between the features of equivalent network depth. In addition, the bottom-up global channel modulation path encodes smaller-scale visual details to a deeper level, enabling the exchange of high-level semantic and low-level detail feature information—ensuring that the fine details of infrared small targets are not overwhelmed by background noise when high-level feature information is exchanged.

3.3. Loss Function

Owing to the problem of severe class imbalance between an infrared image background and typical small targets, we used a soft-IoU [51] loss function during the network training process, which is defined as follows:

$$\ell_{\text{soft-IoU}}(x, s) = \frac{\sum\limits_{i,j} x_{i,j} \cdot s_{i,j}}{\sum\limits_{i,j} s_{i,j} + x_{i,j} - x_{i,j} \cdot s_{i,j}} \tag{10}$$

where $s \in \mathbb{R}^{H \times W}$ is the predicted score map, and $x \in \mathbb{R}^{H \times W}$ is the labelled mask, given in infrared image f.

4. Experimental Evaluation

In this section, we outline quantitative and qualitative evaluations conducted of EAAU-Net on SIRST datasets. Section 4.1 describes our evaluation criteria. The details of the experimental implementation are detailed in Section 4.2. Section 4.3 compares the proposed approach with state-of-the-art methods. The detailed ablation study conducted to verify the efficacy of the components of our proposed network architecture is presented in Section 4.4. Section 4.5 analyses the evaluation results.

4.1. Evaluation Metrics

The signal-to-noise ratio gain (SCRG), background suppression factor (BSF), and receiver operating characteristic (ROC) curve are commonly used as performance metrics for infrared small target detectors. However, we do not consider SCRG and BSF to be suitable in terms of detection performance, as BSF only focuses on the global standard deviation, and SCRG is infinite in most cases. Instead, we use five other metrics—IoU, nIoU, PR curve, ROC curve, and F-area—to further evaluate infrared small target detection methods in the present work. In the formulas below, N is the total number of samples, TP, FP, TN, FN, T, and P denote true positive, false positive, true negative, false negative, true, and positive, respectively.

(1) Intersection-over-union (IoU). IoU is a pixel-level evaluation metric that evaluates the contour description capability of the algorithm. It is calculated as the ratio of the intersection and union regions between predictions and labels, as follows:

$$IoU = \frac{TP}{T + P - TP} \tag{11}$$

(2) Normalised IoU (nIoU). To avoid the impact of the network segmentation of large targets on the evaluation metrics and to better measure the performance of network segmentation of infrared small targets, nIoU is specifically designed for infrared small target detection. It is defined as follows:

$$nIoU = \frac{1}{N} \sum_{i}^{N} \frac{TP[i]}{T[i] + P[i] - TP[i]} \tag{12}$$

(3) PR curve: Precision is used as the vertical axis and recall as the horizontal axis. The closer the curve is to the top right, the better the performance when using the PR curve to show the trade-off between precision and recall for the classifier:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \tag{13}$$

(4) Receiver operating characteristic: The ROC is used to describe the changing relationship between the true positive rate (TPR) and the false positive rate (FPR). They are defined as:

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN} \tag{14}$$

(5) New metric: F-area. F-measure is a precision- and recall-weighted summed average to measure the performance of the harmony. When operating with a fixed threshold, these methods do not sufficiently improve the average accuracy, which is valuable for practical applications. F-area considers both F-measure and average accuracy, taking into account the harmony and potential performance aspects of any technique. It is expressed as given below, where $\beta^2 = 0.3$.

$$F_{measure} = \frac{(\beta^2 + 1) Precision \times Recall}{\beta^2 Precision + Recall} \tag{15}$$

$$\text{F-area} = Average\ Precision\ \times\ F_{measure} \tag{16}$$

4.2. Implementation Details

Datasets. Our proposed EAAU-Net was evaluated on the SIRST datasets, which comprise 427 images and 480 instances with high-quality image annotations [19]. Approximately 55% of these targets occupy only 0.02% of the image area, i.e., only 3 × 3 pixels of the target in a 300 × 300-pixel image. Figure 1 shows some representative and challenging images, from which it may be observed that many targets are very dim, submerged in a complex and cluttered background. In addition, only 35% of the targets in the dataset contain the brightest pixels in the image. In our experiments, we divided the dataset in a 5:2:3 ratio to form a training set, a validation set, and a testing set.

Implementation details. We conducted all CNN- based experiments on the PyTorch platform using a single TITAN RTX GPU, CUDA 10.1, and cuDNN v7. All methods based on traditional manual design were implemented in MATLAB. EAAU-Net was trained using an AdaGrad [52] optimiser, and we set the initial learning rate to 0.05, the batch size to 8, and the weight decay to 1×10^{-4}. The input images were randomly cropped to 480 × 480 pixels, and the network was trained for a total of 300 epochs.

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

To demonstrate the superiority of EAAU-Net, we performed quantitative and qualitative comparisons on the SIRST dataset and compared the proposed network with the state-of-the-art methods top-hat filter [53], max-median filter [5], relative local contrast method (RLCM) [54], multi-scale patch-based contrast measure (MPCM) [55], multiscale grey difference weighted image entropy (MGDWE) [56], local intensity and gradient properties (LIGP) [57], facet kernel and random walker (FKRW) [58], infrared patch-image model (IPI) [10], and reweighted infrared patch-tensor model (RIPT) [11]. These methods are listed in Table 2 with their hyperparameter settings. The CNN-based methods FPN [43], U-Net [31], TBC-Net [25], ACM-FPN [19], ACM-U-Net [19], and ALCNet [24] were also considered.

Table 2. Detailed hyper-parameter settings of traditional methods.

Methods	Hyper-Parameter Settings
Top-hat	Patch size = 3 × 3
Max-median	Patch size $_{\text{median}}$ = 3 × 3
RLCM	Sub-block size = 8 × 8, sliding step = 4, threshold factor k = 1.
MPCM	$N = 1, 3, \cdots, 9$
MGDWE	r = 2, window size = 7 × 7
LIGP	k = 0.2, window size = 11 × 11
FKRW	K = 4, p = 6, β = 200, window size = 11 × 11
IPI	Patch size = 50 × 50, stride = 10, $\lambda = \frac{L}{\sqrt{\min(m,n)}}$, L = 4.5, threshold factor k = 10, $\varepsilon = 10^{-7}$
RIPT	Patch size = 50 × 50, stride = 10, $\lambda = \frac{L}{\sqrt{\min(I,J,P)}}$, L = 0.001, $h = 0.1, \in= 0.01, \varepsilon = 10^{-7}, k = 10$

(1) Quantitative results: For all traditional methods, to obtain better detection performance, we first obtained their predicted values and, then, performed noise suppression by setting a threshold to remove low response areas. The adaptive threshold was calculated by Equation (17). For the CNN-based methods, we used the same experimental parameter settings as in the original works.

$$T_{adaptive} = 0.5avg(G) + 0.5Max(G) \tag{17}$$

where $avg(G)$ and $Max(G)$ denote the average value and largest value of the output, respectively.

Table 3 details the quantitative results of IoU, nIoU, and inference speed for all the methods, and Figure 4 shows a visual comparison in terms of IoU and nIoU; EAAU-Net achieved the best performance in terms of both IoU and nIoU. The significant increase in these values demonstrates that our proposed algorithm significantly improved in terms of accuracy for the shape-matching of prior infrared small targets. The SIRST dataset contains challenging images with different signal-to-noise ratios, background clutter, target shapes, and target sizes; this suggests that our proposed method can learn distinguishing features that are robust to scene variations. Furthermore, as illustrated in Figure 4, the deep-learning-based algorithm achieved a significant improvement over methods based on handcrafted aspects, which are traditionally designed for specific scenes (e.g., specific target sizes and clutter backgrounds), and manually selected parameters (e.g., structure size in top-hat filters and block size in IPI) limit the generalisation performance of these methods.

Table 3. Comparison with state-of-the-art methods on IoU and nIoU.

Methods	IoU	nIoU	Time on CPU/s	Para (M)	Methods	IoU	nIoU	Time on CPU/s	Para (M)
Top-hat	0.295	0.433	0.006	—	RIPT	0.146	0.245	6.398	—
Max-median	0.135	0.257	0.007	—	FPN	0.721	0.704	0.075	1.6
RLCM	0.281	0.346	6.850	—	U-Net	0.736	0.723	0.144	2.2
MPCM	0.357	0.445	0.347	—	TBC-Net	0.734	0.713	0.049	6.93
MGDWE	0.163	0.229	1.670	—	ACM-FPN	0.736	0.722	0.067	1.6
LIGP	0.295	0.410	0.877	—	ACM-U-Net	0.745	0.727	0.156	2.2
FKRW	0.268	0.339	0.399	—	ALCNet	0.757	0.728	0.378	1.44
IPI	0.466	0.607	11.699	—	Ours	0.771	0.746	0.179	2.07

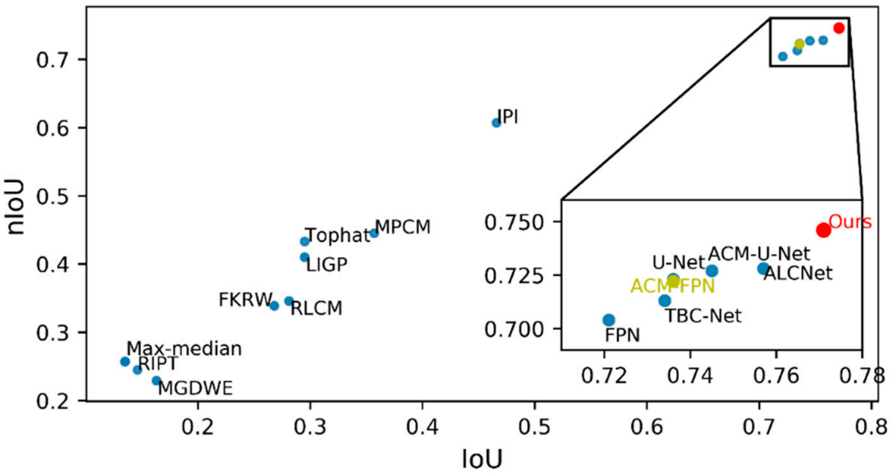
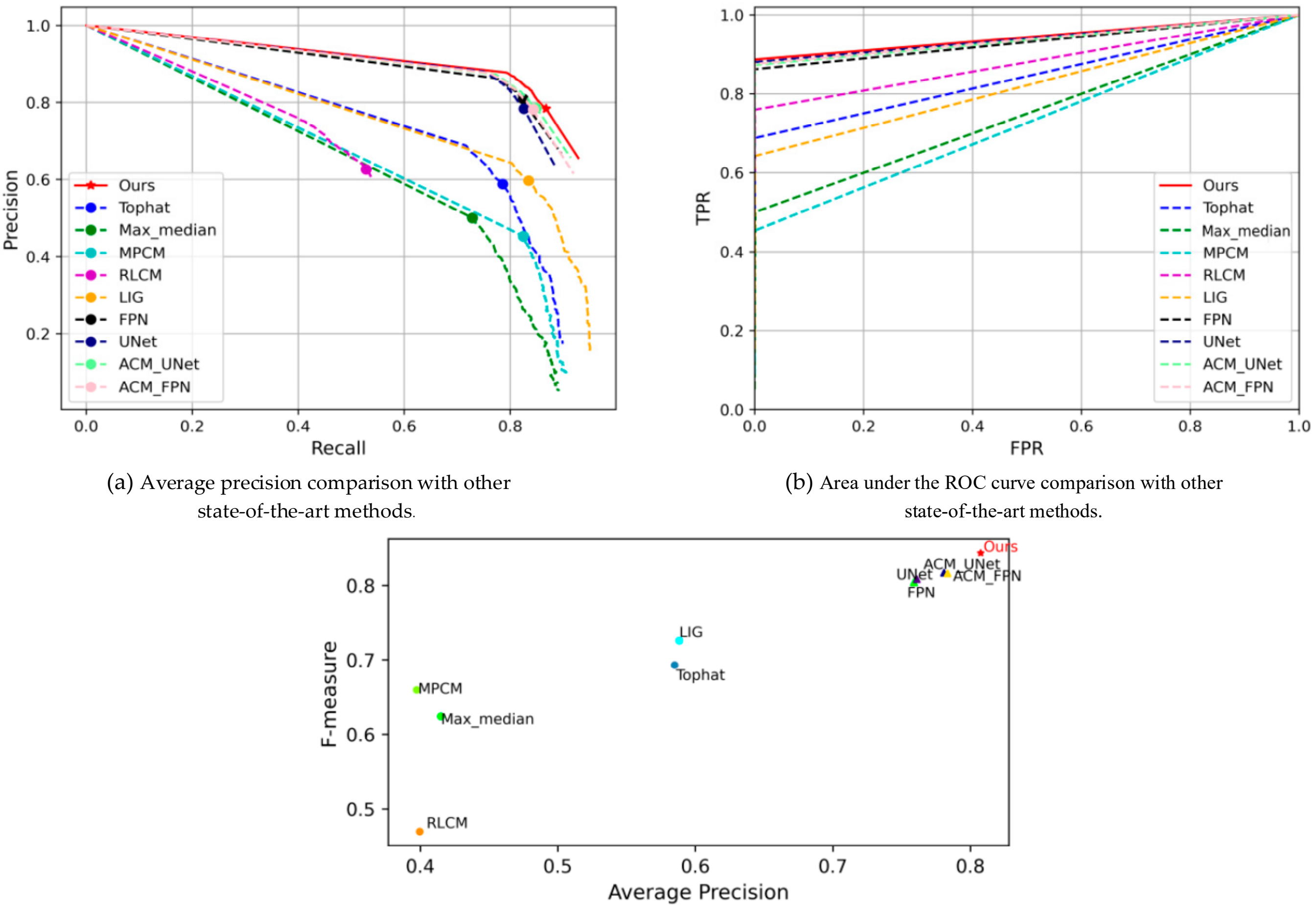


Figure 4. Inference IoU and nIoU comparison on the SIRST dataset.

As presented in Table 3, the improvements achieved by EAAU-Net over other CNN-based approaches (i.e., FPN, U-Net, TBC-Net, ACM-FPN, ACM-U-Net, and ALCNet) are evident. EAAU-Net performed the best; IoU was improved by 0.014, from 0.757 to 0.771, and nIoU by 0.018, from 0.728 to 0.746 for EAAU-Net compared to the next best method, ALCNet, with an increase in network parameters of only 0.63 M. This can be attributed to the design of the new backbone network tailored for infrared small target detection. The U-shaped basic backbone allows for feature fusion across layers through a skip connection and an encoder–decoder structure capable of maintaining and fully learning the primitive features of infrared small targets in the network. In the skip connection and up-sampling paths, we designed the EAA module specifically for infrared small target detection. This module first exchanges information between the channels and spaces by shuffling the feature maps from deep and shallow layers through a shuffle unit. Then, through a bottom-up global channel attention module, fine features from the lower layers are used to dynamically weight and modulate the higher-layer feature maps containing rich semantic information. The EAA module helps the model learn to distinguish features and selectively augment informative features in the deeper layers of the CNN for better performance, significantly improving the detection performance.

Figure 5 illustrates the PR curves, ROC curves, and F-area performance evaluation results against all the state-of-the-art methods. EAAU-Net outperformed all existing CNN-based and traditional handcraft-based infrared small target detection methods for every metric. The PR curve (i.e., Figure 5a) shows that our proposed method was able to achieve the best accuracy and completeness rates, implying that our network has the capacity to focus on the overall target localisation in challenging scenarios where the targets vary in size, type, and location while ensuring detection accuracy. The experimental results for the ROC curve (i.e., Figure 5b) show that our method was able to consistently achieve state-of-the-art performance when the false alarm rate (FA) changed and the probability detection (PD) was able to respond quickly to changes in this false alarm rate. The experimental results for the F-area (i.e., Figure 5c) show that our method still achieved the best performance when both the harmony and accuracy of the algorithm were considered, implying the high potential of our method for practical applications.



(a) Average precision comparison with other state-of-the-art methods.

(b) Area under the ROC curve comparison with other state-of-the-art methods.

(c) F-area comparison with other state-of-the-art methods.

Figure 5. Performance of state-of-the-art methods on the SIRST dataset: (**a**) PR curve, (**b**) area under the ROC curve, and (**c**) F-area.

(2) Qualitative results: Figure 6 shows the 3D visualisation qualitative results of the traditional handcraft-based method and the CNN-based method. Compared with the traditional method, our proposed method was able to produce accurate target localisation and shape segmentation output at a very low false alarm rate. The traditional handcrafted setting-based method is prone to producing several false alarms and missed regions in complex scenes (as shown in Figure 6a,b), owing to the fact that the performance of traditional methods relies heavily on manually extracted features and cannot adapt to changes in the target size. CNN-based methods (i.e., U-Net and ACM-FPN) perform much better than traditional methods. However, U-Net also appears to lose targets (as shown in Figure 6e). EAAU-Net is more robust to these scenario changes. In addition, EAAU-Net was able to generate better shape segmentation than ACM-FPN. This is because we designed the EAA module to help the network adapt well to various types of background clutter, target shapes, and target size challenges using bottom-up cross-layer feature fusion and exchange of channel and spatial information between the same layers, resulting in better performance.

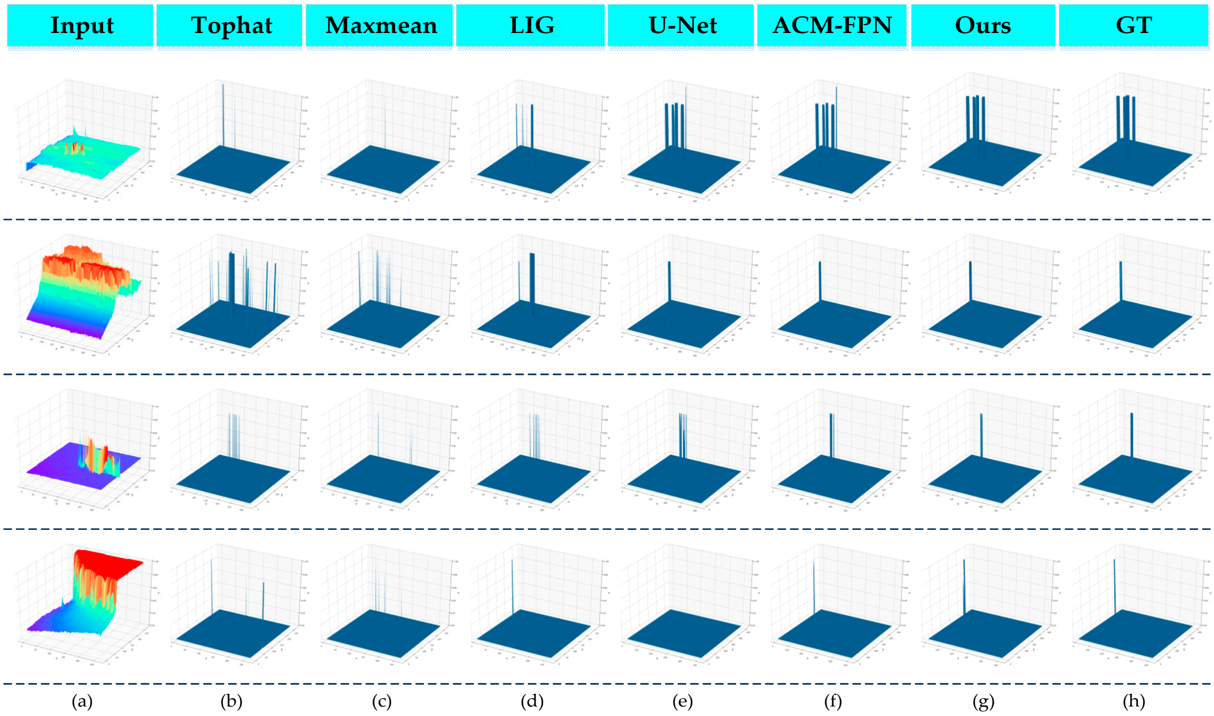


Figure 6. 3D visualisation results of different methods on four test images.

4.4. Ablation Study

In this subsection, we compare the EAA module with several other variants to investigate the potential benefits of our proposed network module and design choices to ensure that the contribution of our proposed model components is justified.

Ablation study for down-sampling depth. The feature information of infrared small targets tends to be very weak, and thus, methods to retain and highlight the deeper features of such targets are of primary concern in target detection network design. Therefore, to avoid losing the deep features of infrared small targets, we applied different down-sampling schemes by changing the block number b in each stage to examine the effects of varying the down-sampling depth. The comparative results are shown in Figure 7; it may be observed that the network performance of EAAU-Net increased gradually with the depth of the network, while the increase in network parameters was less.

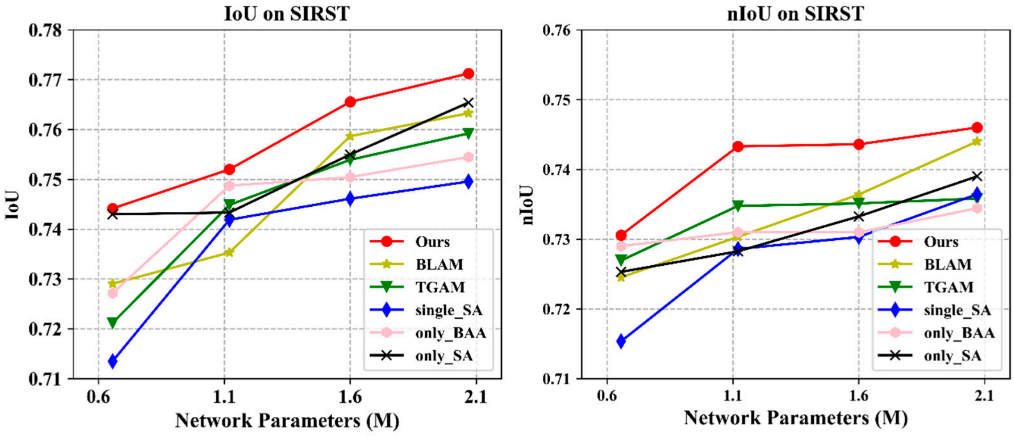


Figure 7. IoU/nIoU comparison with other cross-layer modulation schemes with U-Net as the host networks.

Ablation study for the cross-layer fusion approach. We investigated and compared the bottom-up cross-layer fusion approach with two other ablation modules. The first was a bottom-up local attention modulation (BLAM) module, which aggregates the channel feature context at each spatial location via a local channel attention module (as shown in Figure 8a). The second was a top-down global attention modulation (TGAM) module taking opposite directions from the EAA module and guiding low-level features through high-level coarse information (as shown in Figure 8b). The experimental results are shown in Figure 7, from which it may be observed that the BLAM and TGAM did not perform as well as the EAA module. These results suggest that for infrared small targets, with a given set of computational cost and parameter constraints, fine-grained global contextual information should be aggregated as a guide to refine a high-level feature map by using low-level features as a guide, rather than relying on local information or embedding high-level semantic information into low-level features. Therefore, detailed bottom-up information is more useful for accurate segmentation than top-down semantic guidance.

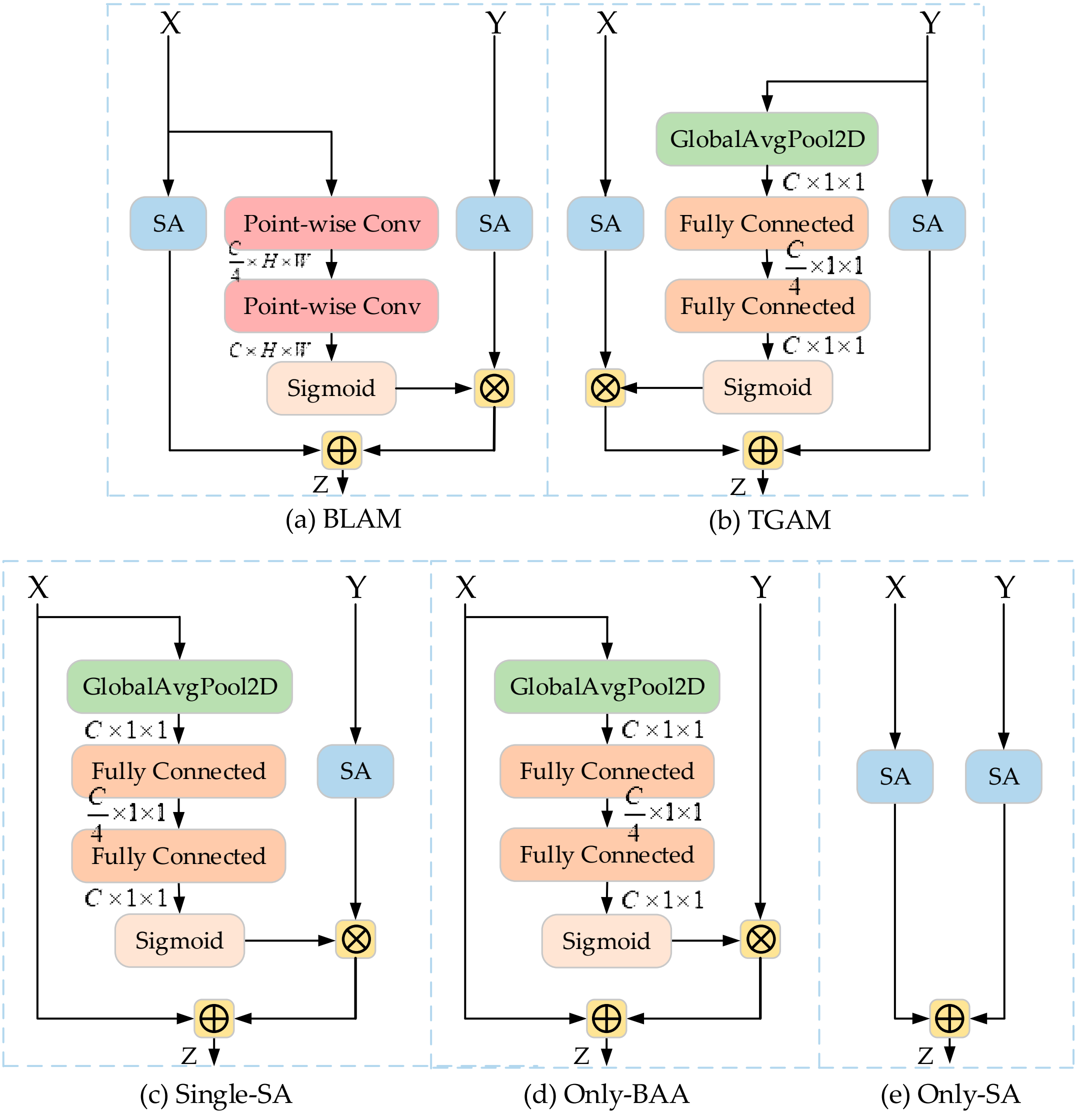


Figure 8. Architectures used in the ablation study on modulation scheme. (**a**) Bottom-up modulation with point-wise channel attention module (BLAM). (**b**) Top-down modulation with global channel attention module (TGAM). (**c**) Bottom-up modulation with single-SA attention module (single-SA). (**d**) Only bottom-up modulation module (only-BAA). (**e**) Only SA block modulation (only-SA). All architectures shared the same hyperparameters.

Ablation study for the SA block. We further investigated and compared the addition of the SA block to both low- and high-level feature maps with two alternate modules. First, a separate SA block considering only spatial and channel information fusion in deep network layers (as shown in Figure 8c) was studied, along with a model lacking an SA block without considering spatial and channel information fusion within the network layers (as shown in Figure 8d), to verify the necessity of adding an SA block to both low- and high-layer networks in the EAA module. Figure 7 provides the results, from which it can be seen that, compared to other modulation schemes, the proposed EAA module performed better in all settings, demonstrating the effectiveness of incorporating an SA block in low- and high-level feature maps to enhance the representation of the CNN by fusing the information of different sub-features using feature dependencies in the spatial and channel dimensions, thus enhancing the original features of the infrared small target.

Ablation study for the EAA module. A comparison between a model using only the BAA block (Figure 8d), SA block (Figure 8e), and the proposed EAA module is given in Figure 7 to verify the effectiveness of the proposed EAA modulation. It can be observed that, compared to using only the BAA block and only the SA block modulation schemes, the proposed EAA module performed better in all settings by exploiting the features in the spatial and channel dimensions and performing information fusion in both low- and high-level feature maps, while using the low-level features obtained by global contextual channel attention in bottom-up pathways to guide the refinement of the high-level feature maps. The results validate the effectiveness of our proposed EAA modulation, i.e., a bottom-up global channel attention mechanism relying on low-level detail information to guide the high-level features for dynamic weighted modulation when low- and high-level feature maps are fused using the spatial and channel feature information. Thus, strong support is provided for the design of bottom-up modulation paths for infrared small target detection and for the fusion of spatial and channel information mechanisms within layers.

4.5. Discussion

Figure 9 shows the qualitative results achieved by the different detection methods on the SIRST dataset. The target areas are enlarged in the lower right corner to allow for a more visual presentation of the fine segmentation results. Correctly detected targets, false positives, and missed regions are indicated by red, yellow, and green dashed circles, respectively. As traditional methods rely on manually extracted features, they cannot adapt to target size and complex background variations, and numerous false alarms and missed detection areas are present. The CNN-based methods (i.e., U-Net and ACM-FPN) are much better than the traditional methods, which also inevitably show varying degrees of false alarms and missed regions due to the infrared small targets being merged into complex backgrounds. EAAU-Net not only generates better shape segmentation for these scene targets, but also generates no false alarms and missed regions, demonstrating the robustness of our proposed method against cluttered backgrounds, dim and small targets, and its better detection performance.

Although our proposed method achieves better performance, it also has the limitation of not being able to accurately segment the boundaries of infrared small targets. Figure 10 shows a partial visualisation of the output of the results of the proposed EAAU-Net on the SIRST dataset. The manually labelled ground truth is ambiguous in terms of one- or two-pixel shifts and had a significant impact on our final IoU and nIoU metrics; for example, a 2 × 2-pixel pinpoint target with even a single pixel of shift would result in the pixel being labelled as 3 × 3; this would result in an error of approximately 50% for that target. Therefore, the proposed EAAU-Net suffers from a certain degree of segmentation error (as shown in Figure 10b,c), which originates from target boundaries that either exceed the labelling mask by a few pixels or segment the target incompletely. Such boundary errors are also present in general vision tasks. Notably, errors are inevitable in manually labelled masks, and the proposed EAAU-Net was able to produce more accurate segmentation results than manually labelled ground truth masks (shown in Figure 10d,e).

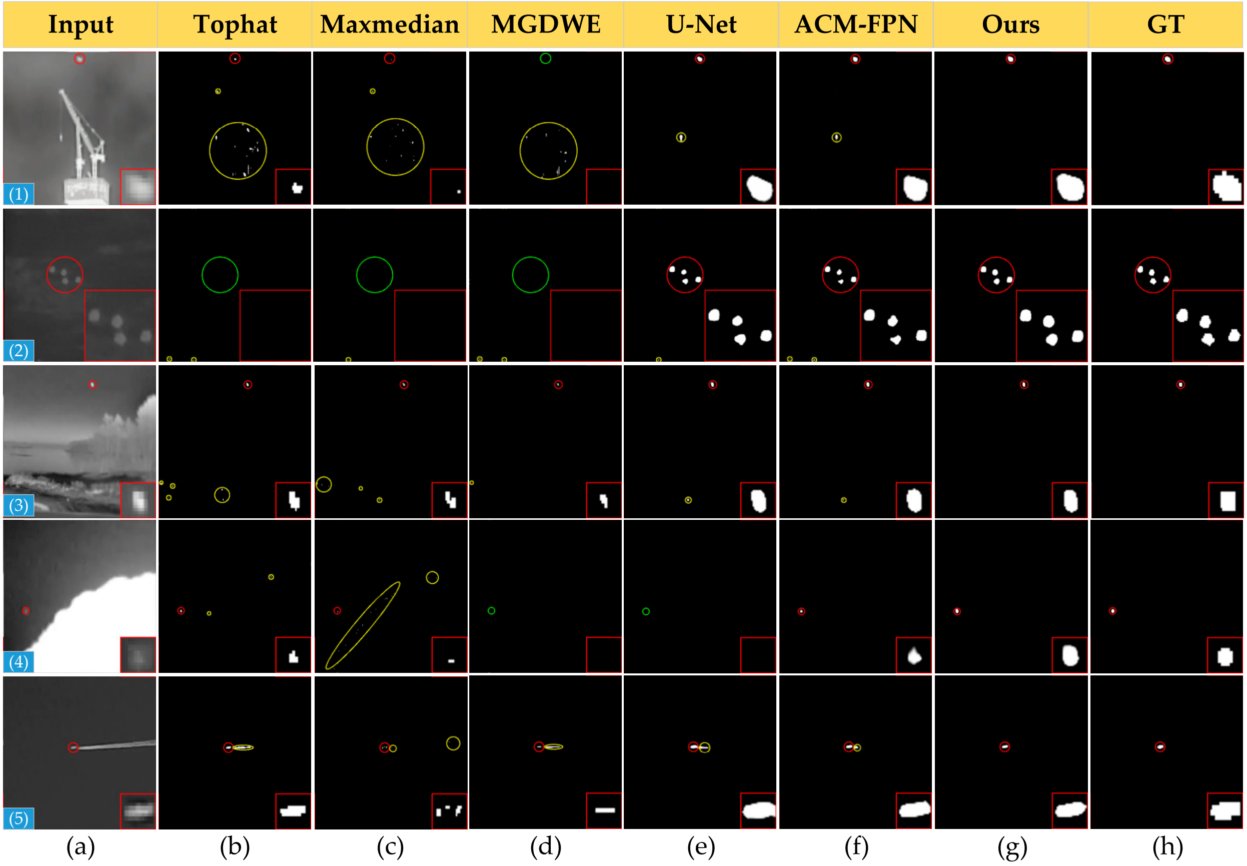


Figure 9. Qualitative results obtained with different infrared target detection methods.

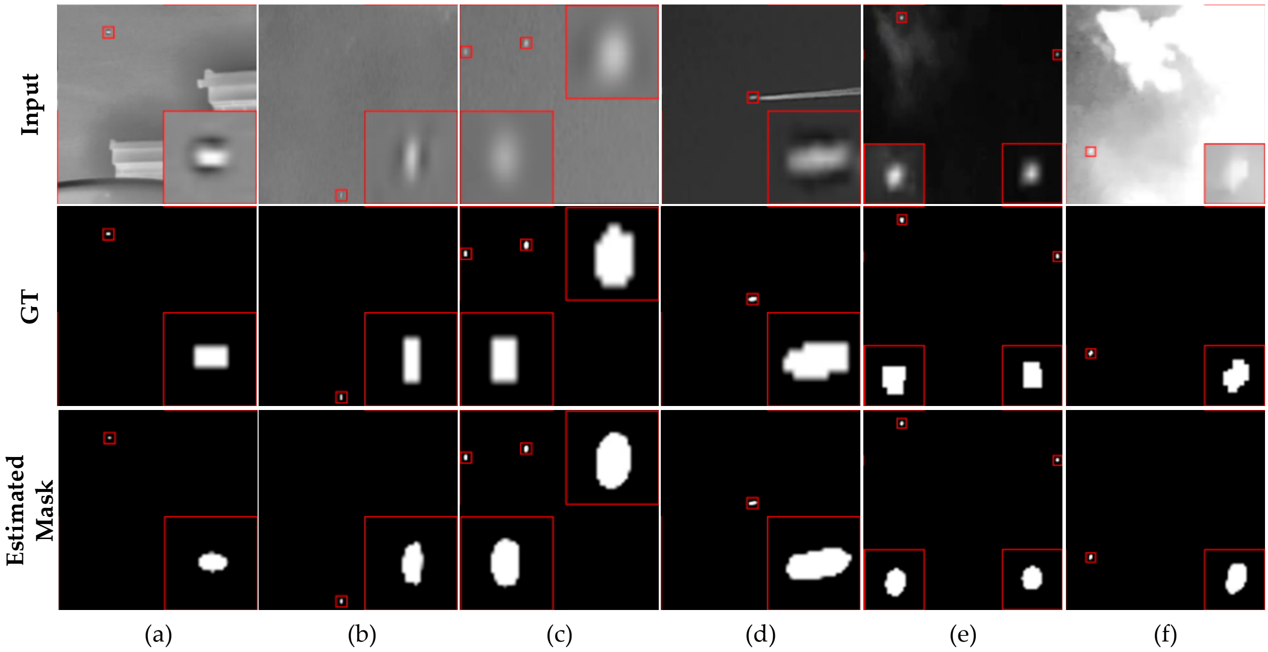


Figure 10. (**a**–**f**) Input image samples, a common ground truth mask (manually labelled), and output of EAAU-Net trained on the SIRST dataset. Our method can even produce more accurate segmentation results than manually labelled ground truth masks.

As can be seen in Figure 10f, the main reason for the inaccurate detection of infrared small targets is that they are too faint. In addition, the small size of the target also results in its small weight in the loss function, which is easily swamped by the boundary error of the larger target during training. In future work, we will focus on reducing the model complexity while adding attention mechanisms and feature fusion modules to the network to enable the exchange of feature information across the different layers of infrared small targets. This is very promising and deserves further research.

5. Conclusions

In this paper, we proposed a lightweight network that fuses contextual attention mechanisms for infrared small target detection. The experimental results show that the network demonstrates strong small target detection capabilities. To retain and highlight the infrared small target features in different layers, the proposed network extracts and fuses the target feature maps in two stages (i.e., in the same layer and across layers), explicitly solving the problem of small targets being lost at deeper layers. With multiple fusions and enhancements, the inherent information of small targets can be fused and fully utilized. Convolutional networks integrating different prior knowledge and deep feature fusion are very promising and deserve further investigation. The EAA module plays an important role in our proposed EAAU-Net by effectively performing spatial and channel feature information exchange within the network layers, while the bottom-up global contextual convolution of low-level feature guidance is used to refine the high-level feature maps, enabling an effective fusion of contextual information to retain and highlight infrared small target features. In addition, we reorganised a set of evaluation metrics to better assess the performance of infrared small target detection algorithms. We conducted an extensive ablation study and compared it with other state-of-the-art methods. The proposed method achieved state-of-the-art results on the publicly available SIRST dataset, suggesting that deep networks should be combined with attention mechanisms, that cross-layer feature fusion schemes preserve targets, and that the adequate information fusion of target features from different layers of the network has the potential to produce better results.

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

Funding: This research received no external funding.

Data Availability Statement: The data used for training and test set SIRST are available at: https://github.com/YimianDai/sirst, accessed on 11 August 2021.

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. Qian, K.; Zhou, H.; Rong, S.; Wang, B.; Cheng, K. Infrared dim-small target tracking via singular value decomposition and improved Kernelized correlation filter. *Infrared Phys. Technol.* **2017**, *82*, 18–27. [CrossRef]
3. Wang, H.; Shi, M.; Li, H. Infrared dim and small target detection based on two-stage U-skip context aggregation network with a missed-detection-and-false-alarm combination loss. *Multimed. Tools Appl.* **2020**, *79*, 35383–35404. [CrossRef]
4. Rivest, J.; Fortin, R. Detection of dim targets in digital infrared imagery by morphological image processing. *Opt. Eng.* **1996**, *35*, 1886–1893. [CrossRef]
5. Deshpande, S.D.; Er, M.; Venkateswarlu, R.; Chan, P. Max-Mean and Max-Median Filters for Detection of Small Targets. In Proceedings of the SPIE's International Symposium on Optical Science, Engineering, and Instrumentation, Denver, CO, USA, 18–23 July 1999. [CrossRef]
6. 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]
7. 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.
8. 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]

9. 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**, *2020*, 1–5. [CrossRef]
10. 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]
11. Dai, Y.; Wu, Y. Reweighted Infrared Patch-Tensor Model With Both Nonlocal and Local Priors for Single-Frame Small Target Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3752–3767. [CrossRef]
12. 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]
13. Zhang, L.; Peng, Z. Infrared Small Target Detection Based on Partial Sum of the Tensor Nuclear Norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
14. Liu, M.; Du, H.; Zhao, Y.; Dong, L.; Hui, M.; Wang, S.X. Image Small Target Detection based on Deep Learning with SNR Controlled Sample Generation. In *Current Trends in Computer Science and Mechanical Automation Vol.1*; Wang, S.X., Ed.; De Gruyter Open Poland: Warsaw, Poland, 2018; pp. 211–220. [CrossRef]
15. 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]
16. 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]
17. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef] [PubMed]
18. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
19. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric Contextual Modulation for Infrared Small Target Detection. *arXiv* **2020**, arXiv:2009.14530.
20. Hou, Q.; Wang, Z.-p.; Tan, F.-j.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust Infrared Small Target Detection Network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *2021*, 1–5.
21. 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**, *2021*, 1–5.
22. 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]
23. Zhang, Q.-L.; Yang, Y. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. *arXiv* **2021**, arXiv:2102.00240.
24. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional Local Contrast Networks for Infrared Small Target Detection. *arXiv* **2020**, arXiv:2012.08573.
25. 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.
26. Qin, Y.; Li, B. Effective Infrared Small Target Detection Utilizing a Novel Local Contrast Method. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1890–1894. [CrossRef]
27. 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]
28. Yang, J.; Gu, Y.; Sun, Z.; Cui, Z. A Small Infrared Target Detection Method Using Adaptive Local Contrast Measurement. In Proceedings of the 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Auckland, New Zealand, 9 September 2019; pp. 1–6.
29. 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]
30. Shi, M.; Wang, H. Infrared Dim and Small Target Detection Based on Denoising Autoencoder Network. *Mob. Netw. Appl.* **2020**, *25*, 1469–1483. [CrossRef]
31. 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: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
32. 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]
33. 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]
34. 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. *IET Image Process* **2021**, *15*, 1–15. [CrossRef]
35. Fu, J.; Liu, J.; Tian, H.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. *arXiv* **2018**, arXiv:1809.02983.
36. Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; Li, S. Sˆ3FD: Single Shot Scale-Invariant Face Detector. *arXiv* **2017**, arXiv:1708.05237.
37. Singh, B.; Najibi, M.; Davis, L. SNIPER: Efficient Multi-Scale Training. *arXiv* **2018**, arXiv:1805.093000.
38. Hu, P.; Ramanan, D. Finding Tiny Faces. *arXiv* **2016**, arXiv:1612.04402.

39. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context Encoding for Semantic Segmentation. *arXiv* **2018**, arXiv:1803.08904.
40. Shrivastava, A.; Gupta, A. Contextual Priming and Feedback for Faster R-CNN. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016. [CrossRef]
41. Fourure, D.; Emonet, R.; Fromont, É.; Muselet, D.; Trémeau, A.; Wolf, C. Residual Conv-Deconv Grid Network for Semantic Segmentation. *arXiv* **2017**, arXiv:1707.07958.
42. Luong, T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.
43. Lin, T.-Y.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. *arXiv* **2016**, arXiv:1612.03144.
44. Hariharan, B.; Arbeláez, P.; Girshick, R.B.; Malik, J. Hypercolumns for object segmentation and fine-grained localization. *arXiv* **2014**, arXiv:1411.5752v2.
45. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1805.10180.
46. Yuan, W.; Wang, S.; Li, X.; Unoki, M.; Wang, W. A Skip Attention Mechanism for Monaural Singing Voice Separation. *IEEE Signal Process. Lett.* **2019**, *26*, 1481–1485. [CrossRef]
47. 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] [PubMed]
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. *arXiv* **2016**, arXiv:1603.05027.
49. Wu, Y.; He, K. Group Normalization. *arXiv* **2018**, arXiv:1803.08494.
50. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. *arXiv* **2018**, arXiv:1807.11164.
51. Rahman, M.A.; Wang, Y. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. In Proceedings of the ISVC 2016: Advances in Visual Computing, Las Vegas, NV, USA, 12–14 December 2016. [CrossRef]
52. Duchi, J.C.; Hazan, E.; Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization; In Proceedings of COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, 27–29 June 2010.
53. 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]
54. 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]
55. 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]
56. 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]
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. 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]