remote sensing

MDPI

Article

IRSTFormer: A Hierarchical Vision Transformer for Infrared Small Target Detection

Gao Chen, Weihua Wang * and Sirui Tan

National Key Laboratory of Science and Technology on Automatic Target Recognition, Collage of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China; chengao18a@nudt.edu.cn (G.C.); tansirui@outlook.com (S.T.)
* Correspondence: wangweihua@nudt.edu.cn

Abstract: Infrared small target detection occupies an important position in the infrared search and track system. The most common size of infrared images has developed to 640 × 512. The field-of-view (FOV) also increases significantly. As the result, there is more interference that hinders the detection of small targets in the image. However, the traditional model-driven methods do not have the capability of feature learning, resulting in poor adaptability to various scenes. Owing to the locality of convolution kernels, recent convolutional neural networks (CNN) cannot model the long-range dependency in the image to suppress false alarms. In this paper, we propose a hierarchical vision transformer-based method for infrared small target detection in larger size and FOV images of 640 × 512. Specifically, we design a hierarchical overlapped small patch transformer (HOSPT), instead of the CNN, to encode multi-scale features from the single-frame image. For the decoder, a top-down feature aggregation module (TFAM) is adopted to fuse features from adjacent scales. Furthermore, after analyzing existing loss functions, a simple yet effective combination is exploited to optimize the network convergence. Compared to other state-of-the-art methods, the normalized intersection-over-union (nIoU) on our IRST640 dataset and public SIRST dataset reaches 0.856 and 0.758. The detailed ablation experiments are conducted to validate the effectiveness and reasonability of each component in the method.

Keywords: infrared small target detection; deep learning; self-attention; transformer

Citation: Chen, G.; Wang, W.; Tan, S. IRSTFormer: A Hierarchical Vision Transformer for Infrared Small Target Detection. *Remote Sens.* **2022**, *14*, 3258. https://doi.org/10.3390/rs14143258

Academic Editors: Pawel Rotter, Wojciech Chmiel and Sławomir Mikrut

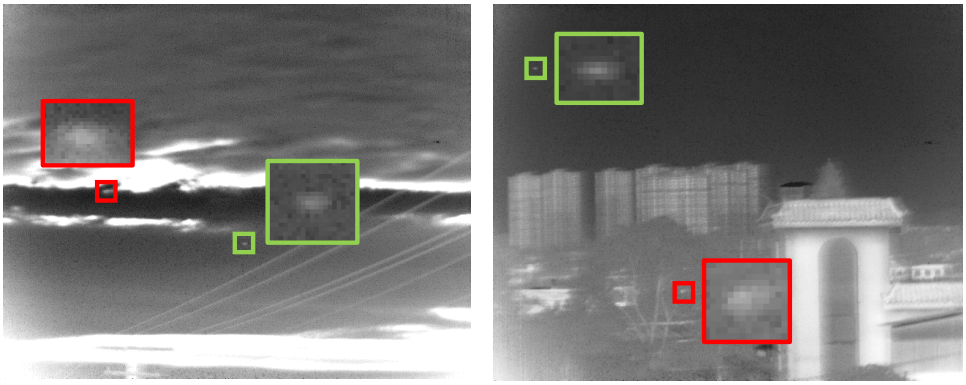
Received: 7 June 2022
Accepted: 4 July 2022
Published: 6 July 2022

1. Introduction

Infrared detectors can detect targets all day long. Even at night, infrared radiation varies depending on the temperature of objects, so the target will have a grayscale difference from its surroundings on the infrared image. In the field of view (FOV) of the infrared search and track (IRST) system, the long-range target contains a small number of pixels, which is generally considered to have an image area of no more than 9 × 9 pixels in the 256 × 256 image according to the definition of SPIE [1]. Detecting infrared small targets is important for military early-warning [2], maritime surveillance [3], etc. However, small scale causes the lack of inherent features of targets, such as shapes, edges, textures, etc. Improving detection rates while reducing false alarm rates has always been a challenging task.

IRST systems usually use multiple frames stitched together to cover a large FOV. The size and FOV of a single-frame image are continuously increased to save time for the wide-area search. Specifically, covering a horizontal 360-degree range, a 320 × 256 infrared detector with the FOV degree of 2.25 × 1.8 requires 200 frames for stitching. Keeping the resolution constant, a 640 × 512 detector with the FOV degree of 4.5 × 3.6 only needs 100 frames. With the same integration time per frame, a detector with the larger size and FOV can save half the total time. However, these improvements bring more interference that can easily be detected as false alarms, making the detection of small targets more

challenging. As shown in Figure 1, in larger size and FOV infrared images, it is difficult to distinguish targets from false alarms from local features only.



(**a**) Sky (**b**) Building

Figure 1. Infrared small targets in developed IRST640 dataset of sky (**a**) and building (**b**) scenes. The green and red box is the target and the false alarm, respectively.

Depending on the images used, infrared small target detection can be classified into single-frame detection and multi-frame detection. In this paper, we focus on single-frame infrared small target (SIRST) detection.

Over the past decades, many methods for SIRST detection have been proposed. Using deep learning as a boundary, these methods can be divided into model-driven and data-driven. Model-driven methods can be further classified as background suppression-based, local contrast measurement (LCM)-based, and optimization-based. Early methods mainly use background suppression, sliding a special window over the image to enhance the target and suppress the background, such as top-hat filter [4], max-median filter [5], etc. However, these methods generate a large number of false alarms when dealing with sea–sky junction lines or heavy cloud clutter. The LCM-based method is inspired by the human visual system, assuming that the target is a local area with a significant grayscale difference from the background. These methods explicitly construct a discriminative measurement that can reflect the characteristics of small targets. The target is detected based on the difference or ratio between the grayscale of the central pixel and the neighboring pixels in the sliding window [6–8]. However, due to the long-range attenuation of infrared radiation and the weak radiation intensity of the target itself, small targets in infrared images often have low grayscale values and do not always satisfy the assumptions of the LCM-based methods. From a matrix perspective, the optimization-based approach models SIRST detection as a low-rank sparse matrix decomposition [9–11]. Ultimately, these model-driven methods are based on hand-designed features and need to follow some specific assumptions. These inherent shortcomings lead to the poor adaptability of these methods to various scenes in increasingly complex infrared images.

Unlike model-driven methods, recent convolutional neural network (CNN)-based methods have the capability of feature learning in a data-driven manner. Publicly available SIRST datasets have further contributed to the development of CNN-based methods [12,13]. Most of these networks consists of a contracting path that extracts high-level features from the input image and an expanding path that reconstructs the mask for pixel-wise segmentation by the single or multilevel up-sampling procedures. In order to detect targets in larger size and FOV images, it becomes critical to learn features of targets and the background in a larger context area. CNN-based methods use the stacking of convolutional layers to increase the receptive field of the network layer by layer, but every value in the feature map only responds to values within the local receptive field in the previous feature map. This inherent locality of convolution makes it difficult to learn long-range dependencies in the image. In NLNet [14], the self-attention mechanism has demonstrated its powerful ability in non-local feature learning in various computer vision tasks and has

been subsequently improved and expanded by other researchers [15,16]. However, these methods never remove the CNN architecture. Self-attention is only used as plug-and-play modules for feature refinement. The transformer is originally used in machine translation to learn the long-range dependency between different word tokens by self-attention layers [17]. ViT first applies it to image classification [18]. Dividing the image into different patches as tokens, the vision transformer for the first time abandons convolution layers to extract image features and instead makes full use of self-attention layers to explicitly learn the long-range dependencies of different patches in the image, opening up a promising direction in the field of computer vision.

Based on the above observation, we purpose a hierarchical vision transformer-based method to detect infrared small targets in larger size and FOV images of 640 × 512, called IRSTFormer. The overall structure inherits the classic encoder-decoder design, learning the pixel-by-pixel segmentation mask in an end-to-end manner for each input image. We design a hierarchical overlapped small patch transformer (HOSPT) to extract multi-scale features from the input image. Image tokens are obtained by the overlapped small patch embedding (OSPE). It also performs down-sampling at different stages to obtain multi-scaled features. In the decoder, we present a top-down feature aggregation module (TFAM), consisting of the multilayer perceptron (MLP) and the channel-attention block. Adjacent feature maps are fused progressively to obtain the final target segmentation mask. Since the pixel share of the infrared small target is extremely small, after analyzing existing binary cross entropy (BCE) and softIoU loss functions, we exploit a simple yet effective combination of them to optimize the network convergence, called combined BCE and softIoU loss (CBS loss). In addition, we develop a publicly available SIRST dataset of 640 × 512, called IRST640, hoping to promote the further development of this field. In summary, the contributions of the paper can be summarized as:

- A hierarchical vision transformer is purposed to detect infrared small targets, which removes the intrinsic shortcomings of existed methods;
- A simple yet effective combination of existing loss functions is exploited to optimize the network convergence;
- Experiments on public SIRST dataset and our developed IRST640 dataset demonstrate the superiority of our method over other state-of-the-art methods.

The rest of the paper is organized as follows. Section 2 reviews related works about infrared small target detection. In Section 3, we describe the proposed method. Sections 4 and 5 is the experiment part, including results and discussion. Section 6 provides the conclusion and our plan about the future work.

2. Related Work

2.1. Detection-Based Infrared Small Target Detection

The data-driven CNN is able to learn features adaptively from images and outperforms model-driven methods for the detection of infrared small targets. According to different processing paradigms, CNN-based methods for SIRST detection can be divided into detection-based [19–22] and segmentation-based methods [12,13,23–29]. The detection-based method outputs the position and scale information of targets directly for the input image, in the same way as generic target detection algorithms, such as Faster RCNN [30] and SSD [31]. ISDet [19] trains both the image filtering network and the target detection network in an end-to-end manner. Du et al. follows the two-stage paradigm of Faster RCNN and designs the small-iou strategy for positive and negative sample partitioning to solve the problem of false convergence and sample misjudgment due to small target size [20]. SSD-ST [21] drops low-resolution layers and enhances high-resolution layers of SSD to adapt the detection of infrared small target. Chen et al. design a two-stage network for target detection in the linear scanning IRST system [22].

2.2. Segmentation-Based Infrared Small Target Detection

The processing paradigm of segmentation-based data-driven methods is the same as that of model-driven methods. They utilize networks to make binary prediction pixel by pixel on the input image to obtain the segmentation mask. After that, a threshold is utilized to output the target position and scale information. Most of the segmentation networks use the encoder-decoder structure, with the encoder condensing the image to extract features and the decoder stretching the features to obtain the segmentation mask. The differences of these methods are reflected in model design [12,23–25], feature optimization [26–29], and feature fusion [13]. Fang et al. converts target segmentation into residual prediction, and the network outputs the background image [23], while training the segmentation network, TBCNet [24] adds a classification network as the semantic constraint to improve the learning ability of the network for image features. MDvsFA [12] and IRSTD-GAN [25] both use the generative adversarial network (GAN) to perform image translation. MDvsFA simultaneously uses two generators to balance missed detection and false alarms. Besides, it publishes a large-scale infrared small target dataset with the image size of 128×128. However, the infrared small targets in this dataset do not exactly meet the definition of SPIE. ACM [13] proposes an asymmetric contextual modulation module to fuse the high-level semantics and low-level details. It also publishes a high-quality dataset called SIRST. This dataset includes various backgrounds with the average size of 302×221, but only contains 427 images. ALCNet [26] modularizes MPCM [32] as special blocks in the network to achieve a learnable local contrast measurement, which improves the feature extraction capability of the network. RISTDnet [27] uses the hand-crafted feature method as convolution layers with fixed parameters and places them at the beginning of the backbone network to form a segmentation network together with normal parameter learnable convolution layers. AGPCNet [28] and LSPM [29] add non-local modules to networks to perform feature refinement. Both of them use the self-attention mechanism to capture long-range dependency in images, but they are still limited by the locality of convolution layers.

2.3. Attention Mechanism

Attention mechanism can be regarded as a resource redistribution mechanism by finding correlations between input data and highlighting important parts. It has been widely used in computer vision and natural language processing. SENet [33] proposes a two-step operation of squeeze and excitation to dynamically modulate the weights of each channel, thus recalibrating the features to improve the representational power of the network. SKNet [34] adds attention branches of different receptive fields on the basis of SENet, realizing adaptive adjustment of receptive fields to the scale of input information. In addition to one-dimension channel attention, CBAM [35] also extracts two-dimension spatial attention from feature maps to re-weight features. NLNet [14] borrows the self-attention mechanism to model the long-distance dependencies in images. To reduce the computational complexity of the non-local module, GCNet [16] proposes a global context modeling module that integrates spatial attention and channel attention into a single module, which can model the global context as efficiently as NLNet and as lightweight as SENet. NLNet has made a preliminary exploration of applicating the self-attention mechanism in deep learning-based computer vision, but like other networks, it only adds plug-and-play modules to the network for feature refinement, without deeply exploring the potential of the attention mechanism. ATAC [36] and AFF [37] utilize attention modules as activation function modules and feature fusion modules in networks, respectively, which essentially suppress useless features and highlight useful features. They can be seen as the exploration of full attention networks.

2.4. Transformer for Computer Vision

With the great success of transformers in machine translation and natural language processing [17], the self-attention mechanism has gradually been applied in computer

vision. As a pioneer, ViT [18] builds a transformer for image classification, which explicitly models the long-range dependencies between different tokens in an image using self-attention layers. For the first time, it abandons the convolution operation in computer vision tasks, thus avoiding the intrinsic locality. After that, the transformer is gradually used for target detection [38], semantic segmentation [39], and super-resolution [40]. In nnformer [41], the transformer is combined with the UNet for medical image processing. MAResU-Net [42] add the self-attention module to CNN for remote sensing image segmentation. After obtaining image features from CNN, Liu et al. adopt the self-attention mechanism to learn the interaction information of image features in a larger range [43]. Unlike it, our network extracts features by a pure transformer structure and does not utilize the convolutional backbone network. For complex infrared images, it is able to explore the long-range dependencies of different regions more effectively and sufficiently.

3. Method

In this section, we introduce our method IRSTFormer in detail, a vision transformer-based method for infrared small target detection.

3.1. Network Architecture

As shown in Figure 2, our method belongs to segmentation-based infrared small target detection. Given an image of $H \times W \times 1$, the network classifies each pixel in the image into target or background, and finally outputs the corresponding segmentation mask.

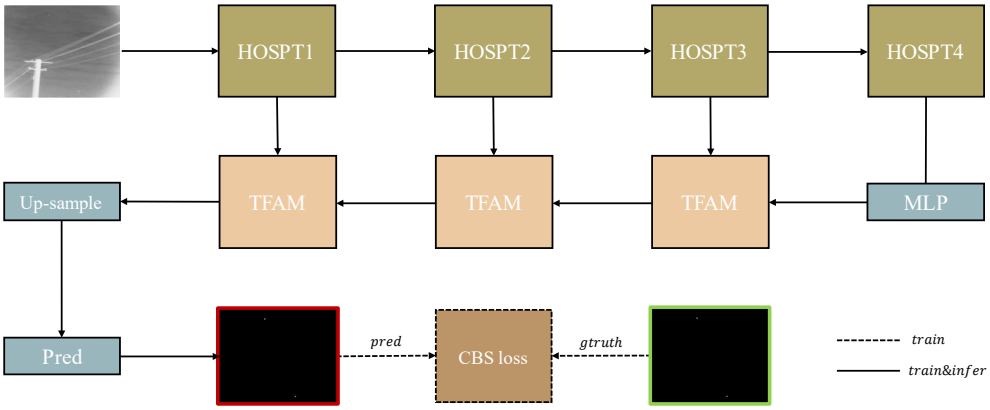


Figure 2. The architecture of the proposed IRSTFormer, which consists of a four-stage encoder hierarchical overlapped small patch transformer (HOSPT), a progressive decoder with top-down feature aggregation modules (TFAM), and the combined BCE and softIoU (CBS) loss.

In order to reduce the false alarm in complex infrared images more efficiently, we propose a hierarchical vision transformer HOSPT to extract multi-scale features of $\frac{H}{2^i} \times \frac{W}{2^i} \times C_i$, where $i \in \{1, 2, 3, 4\}$, $C_1 = 32$, $C_2 = 64$, $C_3 = 160$, $C_4 = 256$. Different from recent CNNs, the self-attention layers in the transformer can learn the dependency relationship in the range of the whole image. This is essential to suppress background interference in complex images. The shallow features contain more target location features that help to locate the target in the image. Deeper features, on the other hand, contain richer semantic features that help to distinguish between false alarms and targets. Therefore, for the decoder, we present the TFAM. In each TFAM, adjacent features are firstly aggregated in the order of top-down. After that, we utilize the channel attention to refine the fused feature. Getting the predicted segmentation mask, we utilize the CBS loss to optimize the network.

3.2. Hierarchical Overlapped Small Patch Transformer

Among the existing deep learning methods, deepening CNNs are used to extract features from infrared small target images, but these methods are always limited by the

locality of convolution, resulting in the poor ability of modeling long-range dependencies in the images. With the increase in size and FOV of infrared detectors, this deficiency is more likely to lead to detection errors. Therefore, we design a transformer-based encoder HOSPT for feature extraction.

At the beginning of each stage, we design the OSPE to dived the input feature map into different patches and conduct linear projection to obtain the two-dimension feature embedding. During this process, the OSPE also completes the down-sampling of feature maps to realize the multi-scale feature extraction. After that, the dot-product self-attention layer can explicitly model the dependencies between different image patches. The extracted features define the importance of how each patch is similar to other patches in the input feature map. Figure 3 shows the structure of every stage.

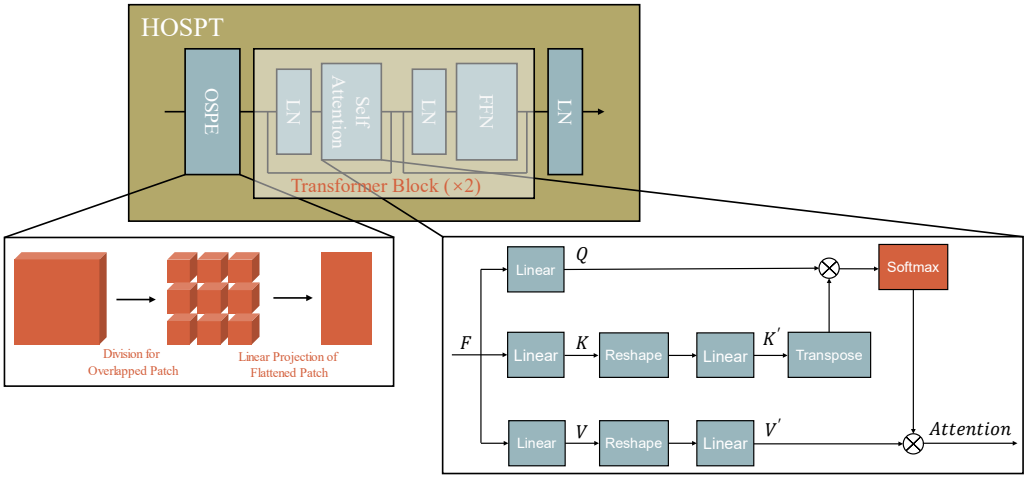


Figure 3. The architecture of each stage in the hierarchical overlapped small patch transformer (HOSPT). The overlapped small patch embedding (OSPE) divides the input feature map into different patches and conducts linear projection to obtain the two-dimension feature embedding. In the self-attention layer, attention features is calculated in form of dot-product. A 3×3 convolution layer is utilized in the feed-forward network (FFN). The layer normalization (LN) is utilized to normalize the feature.

Every stage consists of four parts: OSPE, self-attention layer, feed-forward network (FFN), and layer normalization (LN). One self-attention layer, one FFN, and two LN constitute one transformer block. Each stage consists of two transformer blocks.

After experimenting with different parameters of the OSPE, we set the patch to 3×3 and the stride to 2, which means there is an overlap of three pixels between adjacent patches. Compared with ViT [18], the overlap preserves the continuity between different patches. Specifically, for input three-dimension feature maps of $C_i \times \frac{H}{2^i} \times \frac{W}{2^i}$, it is firstly divided into N_{i+1} patches of $C_i \times 3 \times 3$, where $N_{i+1} = \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$. Then, each patch is flattened and projected linearly into $1 \times C_{i+1}$. Finally, the output two-dimension feature embedding has the size of $N_{i+1} \times C_{i+1}$.

The self-attention layer aims to capture the long-range dependency of every patch pair. As shown in Figure 3, given a feature map, the network learns three sets of parameters to project the features (F) of $N \times C$ to query (Q), key (K), and value (V). Then, the weight is obtained by similarity calculation of the query and the key. Common similarity functions include dot-product, splicing, perceptron, etc. The softmax function is used to normalize the weight. Finally, we multiply the weight with the corresponding value to obtain the final attention features. The attention features define the importance of how each patch is similar to other patches in the feature map. For the original standard multi-head self-attention, it makes Q, K, and V have the same size of $N \times C$ and calculates the self-attention in the form of dot-product with the following equation:

$$Attention = softmax\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V \tag{1}$$

where d_{head} is the dimension. We can see that the computational complexity is quadratic with the size of the feature map, which is prohibitive for large size images. Therefore, the spatial reduction is applied to K and V, which can be formulated as:

$$K^{'} = Linear_{CR\to C}\left(Reshape_{N,C\to\frac{N}{R},CR}(K)\right) \tag{2}$$

$$V^{'} = Linear_{CR\to C}\left(Reshape_{N,C\to\frac{N}{R},CR}(V)\right) \tag{3}$$

K, V of $N \times C$ is firstly reshaped into the size of $\frac{N}{R} \times CR$. Then, the linear projection is utilized to restore the number of channels from CR to C. After such options, we obtain $K^{'}$ and $V^{'}$ of $\frac{N}{R} \times C$. As a result, the computational complexity of self-attention is reduced from $\mathcal{O}(N^2)$ to $\mathcal{O}\left(\frac{N^2}{R}\right)$.

In the FFN, the 3×3 convolution is utilized to replace the position encoding. Therefore, the encoder is robust to different sizes of input images as generally found in the segmentation task. The FFN can be formulated as:

$$x_{out} = MLP(GELU(CONV_{3\times3}(MLP(X_{in})))) \tag{4}$$

3.3. Top-Down Feature Aggregation Module

After obtaining the features of four scales, we should aggregate them in a suitable way. In the U-Net, the transpose convolution and the shortcut are utilized to fuse adjacent scaled features. However, this design will double the number of parameters in the network. Considering the number of parameters in the transformer is already more than that of the CNN, we adopt the simple design of the feature pyramid network (FPN) [44]. In the original FPN, features at different scales are fused by linear addition. This unweighted fusion approach may lead to redundancy of information. Therefore, highlighting important features and suppressing useless features is a more appropriate way to aggregate.

We present the TFAM to form a progressive decoder. As shown in Figure 2, according to the top-down order, the features of adjacent stages are fused to obtain the final pixel segmentation mask. The structure of TFAM is shown in Figure 4, during the fusion, the MLP is first used to unify the dimensions of different scaled features. Then, upper-level features are up-sampled and concatenated with lower-level features along the channel dimension. After that, we utilize a convolution layer of 3×3 and a ReLU function to reduce the dimension and obtain fused features. Finally, channel attention is used to refine the fused features.

Taking features of $C \times H \times W$ as the input, we firstly utilize the global pooling for shrinking the feature maps to obtain channel-wise statistics. Next, channel attention, which explicitly models the global information among channels, is obtained after two linear functions and two activation functions. Refined features can be obtained by multiplying the channel attention and the input features. In this way, useful features can be highlighted while useless features can be suppressed. The overall process can be formulated as:

$$CS = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F(i,j) \tag{5}$$

$$CA = \sigma\left(Linear_{\frac{c}{4}\to c}\left(\delta\left(Linear_{c\to\frac{c}{4}}(CS)\right)\right)\right) \tag{6}$$

$$F_R = F \cdot CA \tag{7}$$

where F means the input features of $C \times H \times W$, CS means the channel-wise statistics, δ means the ReLU function, σ means the sigmoid function, CA means the channel attention, and F_R means the refined features.

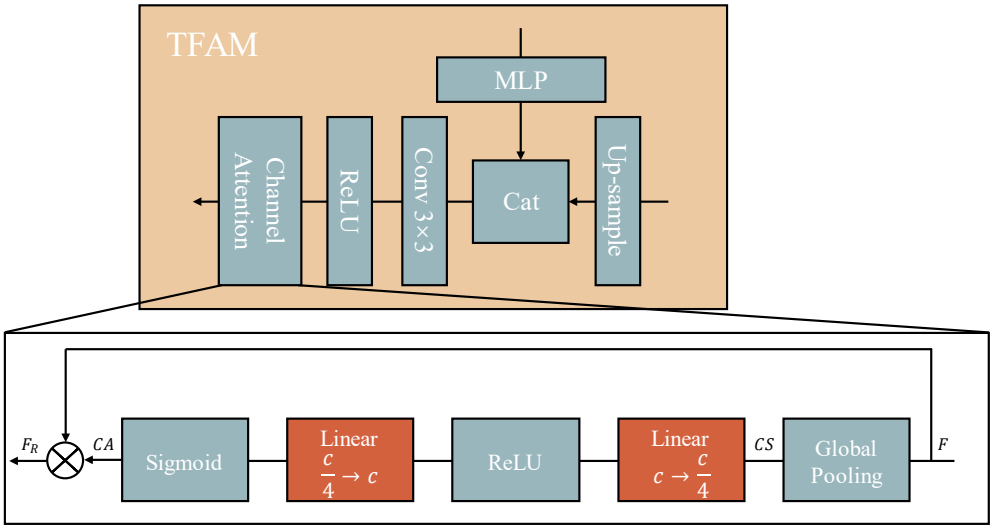


Figure 4. The architecture of the top-down feature aggregation module (TFAM). It mainly consists of the multilayer perceptron (MLP) and the channel-attention block.

3.4. Loss Function

Infrared small target detection can be seen as the binary classification of the input image, where each pixel is distinguished as the target or the background. LSPM [29] utilizes the binary cross-entropy (BCE) loss function when training.

$$L_{BCE} = -\frac{1}{n}\sum_{k=1}^{n}(G_k log P_k + (1 - G_k) log(1 - P_k)) \tag{8}$$

where n is the batch size, G is the ground truth, and P is the predicted segmentation mask. However, the pixel area of small infrared targets is extremely small. In our test images, the small target has a pixel share of less than 0.03% ($\frac{9\times9}{640\times512} \approx 0.00025$). Due to the severe imbalance between positive and negative samples, when training, the network that is supervised by the BCE loss can tend to output zeros because even then the loss function is not very large. In other words, the target is overwhelmed by the background. Secondly, there is no prioritization between the target and background, and all pixels in the image are treated equally. At last, the loss of each pixel is calculated independently, ignoring the global structure of the image.

To obtain a better model, we expect the network to focus more on the target region, rather than treating all pixels equally. Intersection over union (IoU) is usually used as the metric for image segmentation, so an intuitive idea is to directly use IoU as the loss function [45]. In ALCNet [26], AGPCNet [28], and DNANet [46], the softIoU loss function is utilized for infrared small target detection, which is defined as

$$L_{softIoU} = 1 - \frac{1}{n}\sum_{k=1}^{n}\frac{P_k \cap G_k}{P_k \cup G_k} \tag{9}$$

where n is the batch size, G is the ground truth, and P is the predicted segmentation mask. However, when supervised by the softIoU loss, our network cannot converge, resulting in no target can be detected. We analyze this phenomenon from the perspective of the gradient.

For analysis, we assume that the network performs the single point output. Consequently, the following equation is used to calculate the loss value.

$$y = sigmoid(x) = \frac{1}{1+e^{-x}} \tag{10}$$

$$L_{softIoU} = 1 - \frac{t \cdot y + \varepsilon}{t + y - t \cdot y + \varepsilon} = \begin{cases} \frac{y}{y+\varepsilon}, t = 0 \\ \frac{1-y}{1+\varepsilon}, t = 1 \end{cases} \tag{11}$$

where x is the output, $y \in (-1, 1)$ represents the probability value of a pixel being the target, $t \in \{0, 1\}$ represents the ground truth of the pixel, among which 0 means the background and 1 means the target, and ε is the smoothing factor, which is a very small value.

Using the chain rule, the gradient of the softIoU loss is as follows and shown in Figure 5.

$$\frac{\partial L_{softIoU}}{\partial x} = \frac{\partial SoftIoU_{loss}}{\partial y} \cdot \frac{\partial y}{\partial x} = \begin{cases} \frac{\varepsilon y(1-y)}{(y+\varepsilon)^2}, t = 0 \\ \frac{y(y-1)}{1+\varepsilon}, t = 1 \end{cases} \tag{12}$$

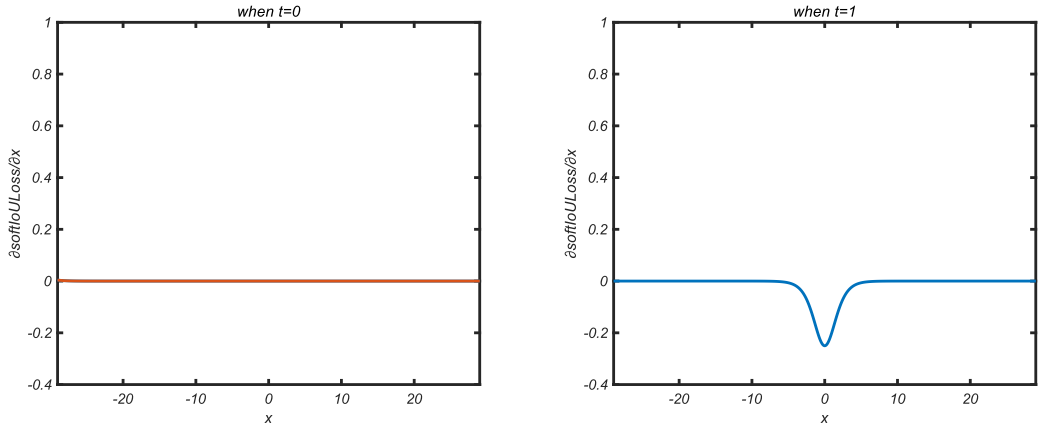


Figure 5. The gradient of the softIoU loss for negative background samples ($t = 0$) and positive target samples ($t = 1$).

Figure 6 shows the network output values x at the first and middle epoch of the training. Because of the weight initialization, the network output values at the first epoch are concentrated around 0. As shown by the gradient diagrams, the absolute values of the gradient at this time are close to 0 for the negative background sample ($t = 0$) and the maximum value for the positive target sample ($t = 1$). This indicates that the contribution of the background region to the network is much smaller than that of the target region, which means that the network is more concerned with finding the target. This tendency will make the network segment more pixels and reduce the missed detection of target pixels. Entering the middle epoch, the network outputs negative and positive values for the predicted background and target regions, respectively. As can be seen from the gradient diagrams, the gradient values at this time both tend to be close to 0. Therefore, the network parameters tend to be updated slowly. Since it has entered the gradient saturation zone, if there are false alarms or missed targets at this time, it will be difficult for the network to overcome these errors, that is, the network is less sensitive to errors.

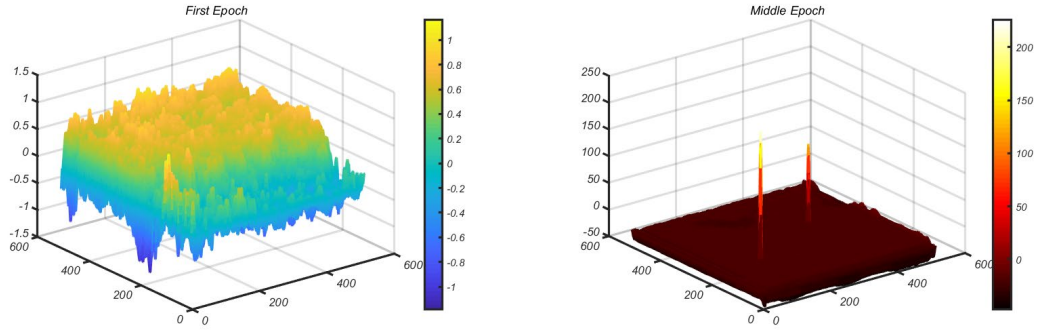


Figure 6. The output x of the network at the first and middle epoch of the training

On the other hand, during the training, the value of the loss function varies with the change of the segmentation mask. To ensure smooth training, large changes in the value need to be avoided. In the softIoU loss, a smaller target pixel area (denominator) leads to a larger change in the loss function values corresponding to the same prediction change (change of the numerator), further leading to a drastic gradient change. Once it enters the saturation zone in the early stage of training, it will make the network difficult to converge. Therefore, compared to the generic instance segmentation, a single softIoU loss leads to an instability of the training when the network performs infrared small target segmentation.

We also analyze the BCE loss function from the perspective of gradient. The following equation is used to calculate the loss value.

$$y = sigmoid(x) = \frac{1}{1+e^{-x}} \tag{13}$$

$$L_{BCE} = -t \cdot \log y - (1-t) \cdot \log(1-y) \tag{14}$$

where x is the output, $y \in (-1, 1)$ represents the probability value of a pixel being the target, $t \in \{0, 1\}$ represents the ground truth of the pixel, among which 0 means the background and 1 means the target. According to the chain rule, the gradient of the BCE loss is as follows.

$$\frac{\partial L_{BCE}}{\partial x} = \frac{\partial L_{BCE}}{\partial y} \cdot \frac{\partial y}{\partial x} = y - t = \begin{cases} y, t = 0 \\ y - 1, t = 1 \end{cases} \tag{15}$$

We can observe that the gradient of the BCE loss is equal to the prediction error. The positive and negative samples contribute to the gradient equally. In the early training period, the prediction error is relatively large. Therefore, the gradient value is large. The network parameters can be updated quickly. Entering the latter period, as the prediction error decreases, parameters are updated slower and the network gradually converges to a stable state.

Based on the analysis above, we propose combined BCE and softIoU (CBS) loss, as formulated below:

$$L_{CBS} = L_{softIoU}(P, G) + \alpha In(1 + \beta \cdot L_{BCE}(P, G)) \tag{16}$$

where $\alpha = 1$, $\beta = 100$. It consists of the BCE loss and the softIoU loss. The former mitigates category imbalance, while the latter can provide smooth gradient values. Inspired by Libra RCNN [47], we utilize the natural logarithm to balance the values of two parts. In Section 5.3, we exploit different forms and parameters of the combination. Compared with the weighted addition, the natural logarithm has the ability to adaptively adjust two loss functions at different epochs of the training. The gradient of the CBS loss is formulated as

$$\frac{\partial L_{CBS}}{\partial x} = \frac{\partial L_{softIoU}}{\partial x} + \frac{\alpha \cdot \beta}{1 + \beta \cdot L_{BCE}} \cdot \frac{\partial L_{BCE}}{\partial x} = \frac{\partial L_{softIoU}}{\partial x} + \frac{100}{1 + 100 \cdot L_{BCE}} \cdot \frac{\partial L_{BCE}}{\partial x} \tag{17}$$

At the early epoch of the training, the value of L_{BCE} is relatively large. Due to adjustment coefficients α and β, the contribution of L_{BCE} to the gradient is reduced, and the gradient mainly comes from $L_{softIoU}$. The network focuses more on the target area. When the training enters late epochs, the value of L_{BCE} becomes small. Due to the effect of adjustment coefficients, the gradient contribution of L_{BCE} is increased. Even if the error occurs, resulting in the saturation of $L_{softIoU}$, L_{BCE} still can supplement enough gradient for network parameters to continue iterating.

4. Result

In this section, we first introduce the experimental setting including the dataset, evaluation metrics, and implementation details. Then, we compare our IRSTFormer with other state-of-the-art methods to demonstrate the effectiveness of the network. Finally, we show ablation studies on the encoder, decoder, loss function, and dataset of the network to verify the design of the network.

4.1. Experimental Setting

4.1.1. Dataset

Our study is motivated by the fact that infrared images with progressively larger size and FOV make small target detection more challenging. However, existing public datasets do not meet this requirement. Collecting by an IRST system, we develop a synthesized infrared small target dataset of 640 × 512, called IRST640, which contains 1024 images. As shown in Figure 1, the background interference includes the cloud, buildings, and trees. We generate one or more infrared small targets on each real scene image. Zhao et al. have demonstrated the potential of synthesized data for realistic detection tasks [25]. The IRST640 is available on our homepage: https://github.com/jzchenriver/IRST640 (accessed on 5 June 2022).

In the ratio of 8:2, we obtain the training set of 819 images and the test set of 205 images. Since our dataset is collected by an IRST system at a fixed location, the images are of a single scene. The publicly available dataset SIRST has an average image size of 302 × 221. Although smaller than ours, it contains more kinds of scenes. After experimenting based on the mixed and unmixed dataset, we mix the training set of our IRST640 with the training set of SIRST to avoid overfitting caused by the single scene.

The final training set consists of 1160 images, among which 819 images come from our IRST640 dataset and 341 images come from the public SIRST dataset. Two test sets of them keep independent. To be specific, 205 images from the IRST640 dataset and 86 images from the SIRST dataset are used for evaluation separately.

4.1.2. Evaluation Metrics

The network performs detection by segmenting targets from the background, so pixel-level metrics and target-level metrics are utilized to conduct the evaluation simultaneously.

In the terms of pixel-level metrics, we use the normalized IoU (*nIoU*) and FMeasure (*FM*) to perform the comprehensive evaluation. They are defined as:

$$nIoU = \frac{1}{n}\sum_{k=1}^{n}\frac{TP_k}{T_k + P_k - TP_k} \tag{18}$$

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

$$FMeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{21}$$

where TP, FP, FN, T, and P denote the true positive, false positive, false negative, true, and positive, respectively. The nIoU first calculates the IoU for each prediction and then calculates the mean value for the entire test set. The FM simultaneously considers Precision and Recall. A method is considered to be a good method when obtaining high values (↑) on all of these metrics.

Target-levels metrics probability of detection (PD) and false-alarm rate (FA) are also utilized. They are defined as:

$$PD = \frac{\#numberofcorrectlypredictedtargets}{\#numberofalltargets} \tag{22}$$

$$FA = \frac{\#numberoffalselypredictedpixels}{\#numberofallpixels} \tag{23}$$

We consider that the correctness of the prediction depends on whether the centroid distance between it and the ground truth is less than 3 pixels. A method is considered to be a good method when obtaining a high PD (↑) and a low FA (↓).

4.1.3. Implementation Details

Our proposed method is implemented using Pytorch 1.7.0. We resize the image to the size of 512 × 512 as the input. The network is optimized by the adagrad method, where the weight decay coefficient is set to 0.01. The pre-trained weight on the ImageNet is used for network initialization. We train the network for 100 epochs with a batch size of 2. In the first 10 epochs, the learning rate increases linearly from 0 to 0.0005. After that, the initial learning rate is multiplied by $\left(1 - \frac{epoch-10}{total_{epoch}-10}\right)^{0.9}$ for every epoch. On the hardware, we implement all methods on a Ubuntu PC with one Nvidia RTX 2080ti GPU and two Intel Xeon E5-2678 CPUs.

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

We perform the comparison to other 13 methods to demonstrate the superiority of our proposed IRSTFormer, including background suppression-based methods (Tophat [4], Max-median [5]), LCM-based methods (TLLCM [6], RLCM [7], AAGD [8]), optimization-based methods (PSTNN [9]), CNN-based methods (ResNetFPN [44], ALCNet [26], DNANet [46], AGPCNet [28], LSPM [29]), GAN-based methods (MDvsFA [12]), and transformer-based methods (Segformer [39]).

4.2.1. Qualitative Results

For the intuitive comparison, Figures 7 and 8 show the segmentation masks of the total of 14 methods corresponding to the same image.

For the complex cloud background shown in Figure 7a, neither the background suppression-based methods nor the optimization-based method detects the target. Although TLLCM and RLCM can detect the target, both are less effective in suppressing false alarms. All the deep learning-based methods are able to accurately detect targets from complex backgrounds, indicating that their feature extraction ability learned from training data is robust to heavy cloud layers.

In Figure 7b–d, the ground scene is more complex than the sky scene. The deep learning-based methods show false alarms. In particular, in Figure 7c, only ResNetFPN, DNANet, and our proposed IRSTFormer accurately detect the target and no false alarms appear.

In Figure 7e, the target luminance is weak. Furthermore, there are a large number of noisy pixels. The detection results of the deep learning-based methods are overall inferior to the traditional methods. MDvsFA detects both the target and the noisy pixels, so the FA value is high. Even so, our proposed method achieves target detection with zero false alarms.

In summary, benefiting from the feature learning capability of networks, the detection results of deep learning-based methods are much better than that of traditional methods.

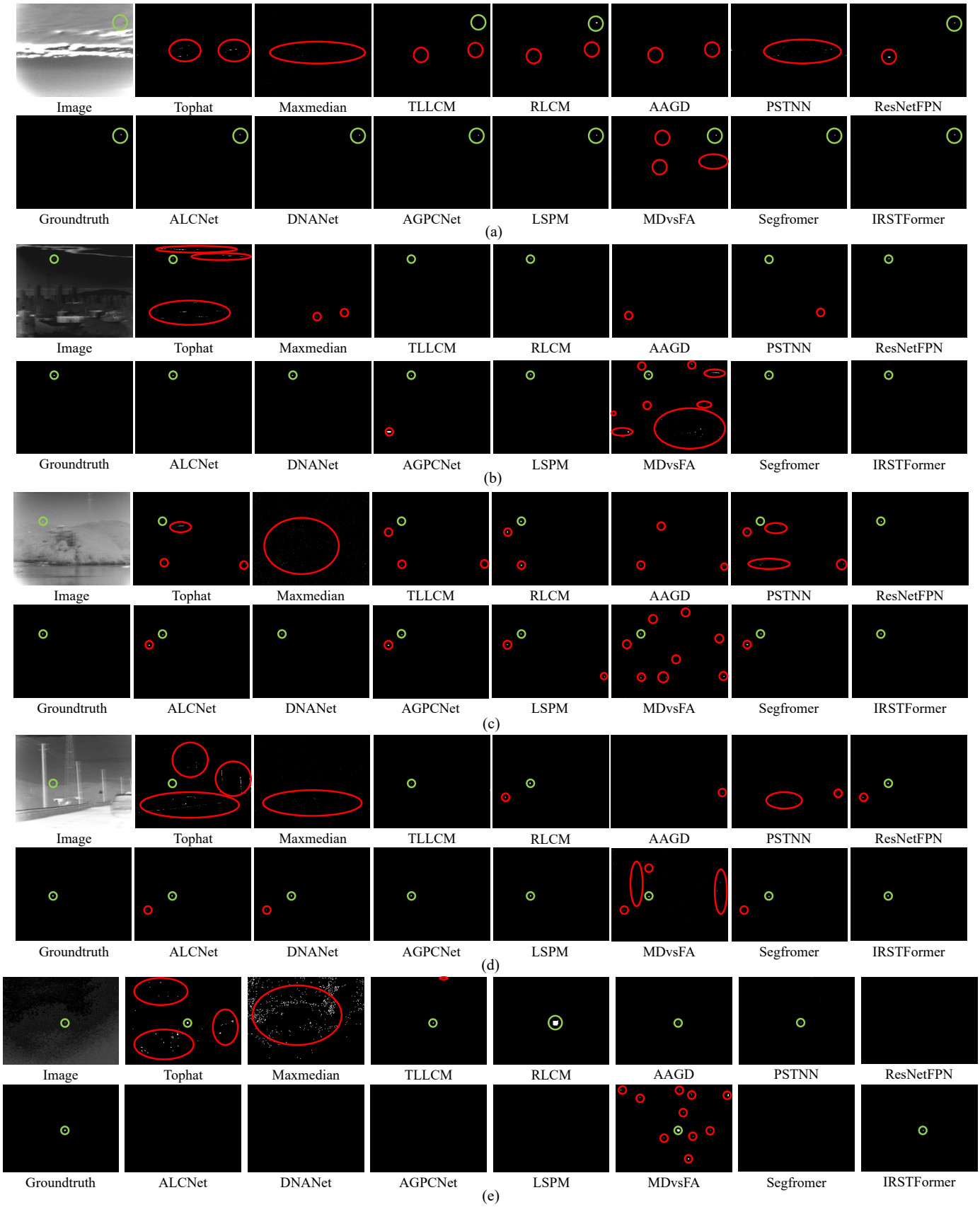


Figure 7. The segmentation masks of 14 methods corresponding to the same image of cloud (**a**), ground (**b**–**d**), and noisy pixels (**e**) scene. The green and red circle is the detected target and false alarm, respectively.

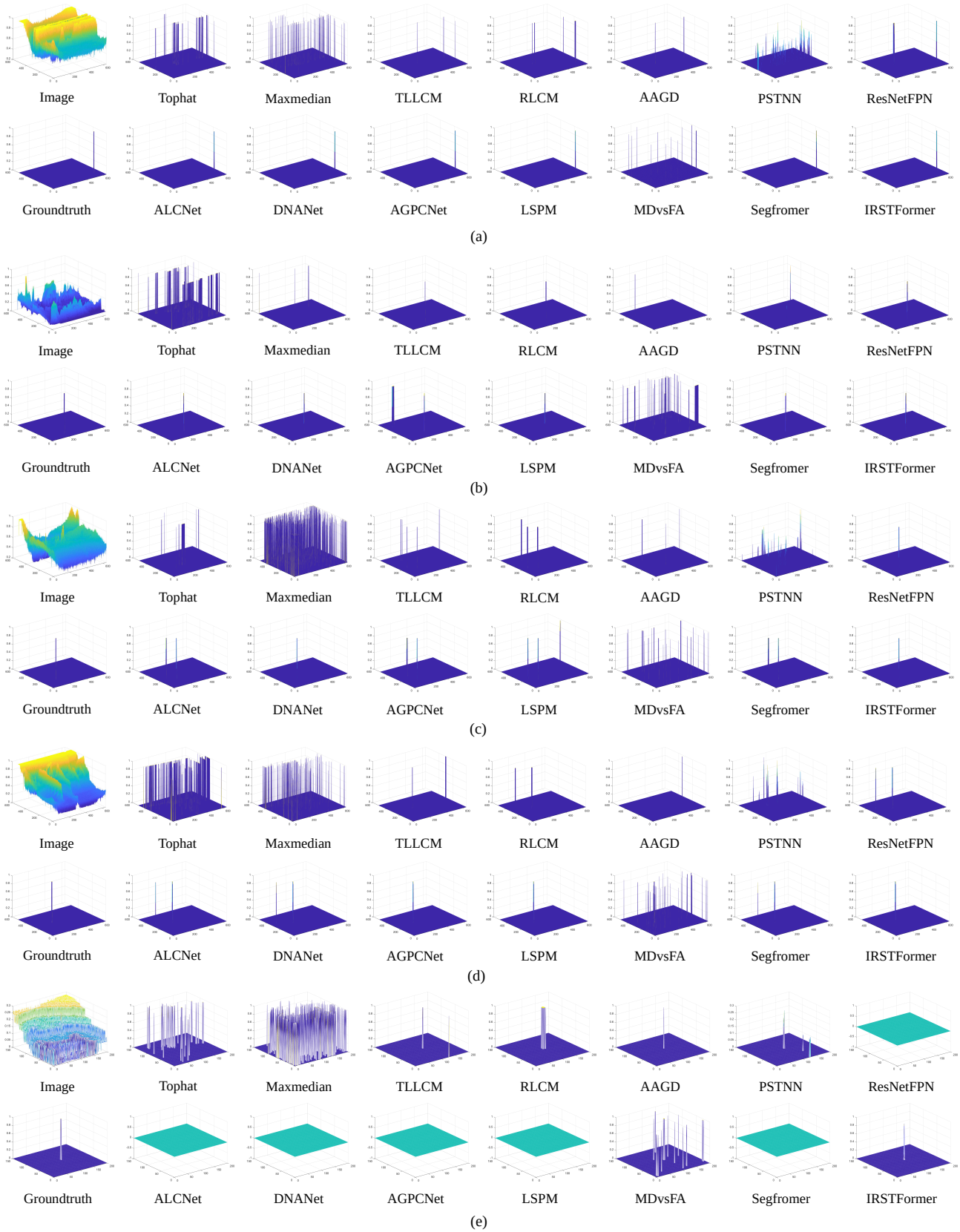


Figure 8. The 3D segmentation masks of 14 methods corresponding to the same image of cloud (**a**), ground (**b**–**d**), and noisy pixels (**e**) scene.

Among the three subclasses: CNN, GAN, and Transformer, the GAN-based MDvsFA has a higher false alarm rate although it is more capable of detecting targets. For cloud scenes with relatively simple features, the CNN-based method has a strong detection capability. However, once faced with more complex ground scenes, it cannot suppress false alarms well due to the locality of convolution. With the increase in the FOV and size of infrared images, this defect causes more detection errors. The transformer-based approaches explicitly model the dependencies between different image patches through the self-attention mechanism, thus having strong ability to distinguish between targets and false alarms. Compared with Segformer, our proposed IRSTFormer has improvements for the encoder, decoder, and loss function, therefore achieving both improvements in detection rate and reduction in false alarm rate.

4.2.2. Quantitative Results

The quantitative results of different methods on IRST640 and SIRST are shown in Table 1.

Table 1. The comparison results on the IRST640 and SIRST dataset with pixel-level and target-level metrics.

Method	IRST640				SIRST			
	$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$	$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$
Tophat	0.0836	0.144	0.324	2000	0.425	0.567	0.587	2000
Max-median	0.0279	0.0516	0.396	3000	0.253	0.382	0.463	412.5
TLLCM	0.241	0.383	0.450	71.06	0.283	0.411	0.424	203.1
RLCM	0.464	0.616	0.896	228.3	0.339	0.470	0.684	2000
AAGD	0.0648	0.116	0.216	135.5	0.165	0.271	0.348	157.0
PSTNN	0.286	0.398	0.671	1000	0.626	0.726	0.785	1000
ResNetFPN	0.806	0.884	0.985	67.51	0.664	0.763	0.872	582.4
ALCNet	0.829	0.903	0.988	6.171	0.684	0.785	0.881	248.8
DNANet	0.808	0.890	0.976	2.047	0.645	0.747	0.844	773.1
AGPCNet	0.761	0.852	0.985	145.2	0.674	0.778	0.908	557.6
LSPM	0.838	0.908	0.988	2.618	0.723	0.825	0.936	303.0
MDvsFA	0.368	0.518	0.918	1000	0.332	0.472	0.906	9000
Segformer	0.761	0.852	0.979	189.6	0.664	0.767	0.899	358.4
IRSTFormer	**0.856**	**0.920**	**0.988**	**1.496**	**0.758**	**0.859**	**0.991**	**57.66**

Our proposed IRSTFormer achieves optimal results on all metrics, which proves the effectiveness of the method. Specifically, deep learning-based methods significantly outperform traditional methods in both pixel-level metrics and target-level metrics. This is because traditional methods cannot learn to extract features from the image. Moreover, many complex scene images in the test set do not satisfy the prior assumptions of traditional methods, such as buildings, trees, clouds, etc. This leads to poor results of these methods, among which RLCM and PSTNN achieve relatively good results.

For deep learning-based methods, the input size is all of 512×512. The PD of all methods exceeds 0.84, but the detection results of GAN-based MDvsFA show a large number of false alarms, which lead to a low level of pixel-level metrics FA. For CNN-based DNANet and AGPCNet, in the experiments, the softIoU loss used in the original paper does not enable the network to converge properly. We analyze the reason for this phenomenon in Section 3.4. After changing the loss function to the proposed CBS loss, the training proceeds normally. It is worth noting that LSPM, borrowing the idea of self-attention, has suboptimal detection results, but it only use the self-attention as feature optimization

modules in the network. Compare with it, we construct a feature extraction network that is entirely based on the self-attention mechanism, taking full advantage of its ability to model long-range dependencies. On the IRST640 dataset, the nIoU and FM are improved by 0.018 and 0.012. On the SIRST dataset, which contains more complex scenes, the most noticeable improvement is on the PD metric, which increases by 0.055. Compared with the transformer-based Segformer, the improvement acquired by our method in both nIoU and FM is close to 0.1. PD continues to improve while FA drops by an order of magnitude.

In Table 2, based on a computer with one Nvidia RTX 2080ti GPU, we show the computational complexity (GMac), the number of parameters, and the processing speed (FPS) of different networks when a 512×512 image is input. Compared with the suboptimal LSPM, our method achieves an improvement in all experimental metrics, while the computational complexity and the number of network parameters decrease significantly. In the future work, we will focus on how to improve the processing speed of the IRSTFomer.

Table 2. The computational complexity, the number of parameters, and the processing speed of different networks.

Method	Complexity (GMac)	Parameters	Speed (FPS)
ResNetFPN	15.28	374.56 K	119
ALCNet	14.52	384.79 K	53
DNANet	56.41	4.70 M	11
AGPCNet	172.54	12.36 M	7
LSPM	246.25	31.58 M	30
MDvsFA	988.44	3.77 M	5
Segformer	6.74	3.71 M	65
IRSTFormer	111.06	4.82 M	22

5. Discussion

As mentioned earlier, our proposed IRSTFormer consists of three parts: HOSPT encoder, TFAM decoder, and CBS loss. In this section, we discuss our method in detail. Each experiment is trained for 50 epochs.

5.1. Ablation Study

Using Segformer with BCE loss as the baseline, we show the enhancement of each of these three parts. Tables 3 and 4 show the results on two datasets.

Table 3. The ablation study on the IRST640 dataset.

HOSPT	TFAM	CBS Loss	IRST640			
			$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$
			0.761	0.852	0.979	189.6
✓			0.798	0.882	0.988	27.30
✓	✓		0.828	0.902	0.991	16.08
✓	✓	✓	0.856	0.92	0.991	1.496

Table 4. The ablation study on the SIRST dataset.

HOSPT	TFAM	CBS Loss	SIRST			
			$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$
			0.664	0.767	0.899	358.4
✓			0.728	0.836	0.972	571.3
✓	✓		0.731	0.836	0.963	270.6
✓	✓	✓	0.743	0.845	0.982	124.6

On the IRST640 dataset, after changing the encoder into the HOSPT, nIoU is improved by 0.037. It indicates that the design of overlapped small patches is suitable for segmenting the infrared target on an extremely small scale. Based on this, the TFAM and CBS loss continue to improve the detection performance. Especially, the CBS loss is highly effective for false alarm suppression. Compared to the baseline, our method acquires an increase of 0.095 in nIoU, 0.068 in FM, and 0.012 in PD. Furthermore, FA drops by more than 100 times. The performance on the SIRST dataset is similar. The target-level metric PD is improved by nearly 10% and FA decreases by nearly 3 times. Three parts improve the nIoU by 0.064, 0.003, and 0.012, respectively, which demonstrates the effectiveness of OSPE in the transformer encoder. Another pixel-level metric FM shows the same trend. In the next section, we show the influence of different parameters in the OSPE.

5.2. Different Parameters in the OSPE

In this section, we experiment with different parameters in the OSPE. The results are shown in Tables 5 and 6.

Table 5. Results of different parameters in the OSPE on the IRST640 dataset.

Patch Size	Stride	IRST640			
		$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$
7	2	0.798	0.879	0.979	28.05
5	2	0.796	0.877	0.985	41.51
3	2	0.798	0.882	0.988	27.30
2	2	0.766	0.860	0.994	70.87

Table 6. Results of different parameters in the OSPE on the SIRST dataset.

Patch Size	Stride	SIRST			
		$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$
7	2	0.670	0.777	0.927	320.7
5	2	0.703	0.811	0.972	255.5
3	2	0.728	0.836	0.972	571.3
2	2	0.700	0.807	0.936	749.2

As mentioned above, the HOPST has four stages. The OSPE is located at the beginning of each stage. It divides the feature maps into patches and projects them to the two-dimension embedding. There are two parameters: patch size and stride size, among which the stride size determines the down-sampling times. Considering the target size, we set the stride to 2. Therefore, the downsampling times of the last stage is 16. It is worth noting that, compared with the Segformer that downsamples 32 times, this small change results in a significant improvement in detection metrics. The experiments on both datasets show

the same trend. When the size of the patch drops from 7 to 3, the detection performance gradually improves. We suggest that this is related to the pixel size of infrared small targets in the images. When the patch and stride size is 2 and 2, it means there are no overlapping pixels between adjacent patches. From the result, we can see that the presence of overlapping pixels leads to better results. This indicates the importance of preserving the continuity between patches.

5.3. Different Forms of Combination of the BCE and SoftIoU Loss

After analyzing the existing loss functions for infrared small target detection, we believe that both BCE loss and softIoU loss have their own drawbacks. When the size and FOV of infrared images increase, these drawbacks will lead to more detection errors. We exploit the suitable form to combine them. The simplest form is the weighted addition (WA), as formulated below,

$$L_{softIoU} + \alpha \cdot L_{BCE} \tag{24}$$

In addition, inspired by LibraRCNN [47], we also use the natural logarithm (NL) to balance the values of two parts, which is formulated as,

$$L_{softIoU} + \alpha \ln(1 + \beta \cdot L_{BCE}) \tag{25}$$

The results are shown in Tables 7 and 8. The trend of the experimental results on the two test sets is similar, with the best performance coming from the last set in the table—two parts are combined by the natural logarithm with $\alpha = 1$, $\beta = 100$. When the softIoU loss is utilized, the network cannot converge during the training, and all metrics are zero. We analyze the cause of this phenomenon in Section 3.4. For the weighted addition and natural logarithm, the latter group has better results. We suggest that the natural logarithm function has the ability to adaptively adjust the trend of the loss value at different epochs of training. Meanwhile, the nonlinearity of the function improves the ability of the network to fit nonlinear data. Because the values of two functions differ by many orders of magnitude, in the respective groups, the best results simultaneously come from $\alpha = 1$, $\beta = 100$.

Table 7. Results of different forms of loss function combination on the IRST640 dataset.

Form	α	β	IRST640			
			$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$
BCE	-	-	0.828	0.902	0.991	16.08
softIoU	-	-	0	0	0	0
WA	1	-	0.837	0.908	0.991	15.33
WA	10	-	0.845	0.912	0.991	22.81
WA	100	-	0.849	0.916	0.991	3.927
NL	1	1	0.823	0.898	0.985	38.15
NL	1	10	0.85	0.917	0.991	23.46
NL	1	100	0.856	0.92	0.991	1.496

Table 8. Results of different forms of loss function combination on the SIRST dataset.

Form	α	β	SIRST			
			$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$
BCE	-	-	0.731	0.836	0.963	270.6
softIoU	-	-	0	0	0	0
WA	1	-	0.671	0.777	0.963	834.4
WA	10	-	0.698	0.801	0.945	407.6
WA	100	-	0.727	0.826	0.972	158.4
NL	1	1	0.634	0.735	0.853	359.7
NL	1	10	0.702	0.803	0.908	236.0
NL	1	100	0.743	0.845	0.982	124.6

5.4. Different Training Sets

We conduct experiments on two datasets: our proposed IRST640 and publicly available SIRST. Table 9 shows the results with different training datasets.

Table 9. Results of different training sets.

Training	Test: IRST640				Test: SIRST			
	$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$	$nIoU \uparrow$	$FM \uparrow$	$PD \uparrow$	$FA \downarrow (\times 10^{-7})$
IRST640	0.858	0.921	0.991	0.3722	0.393	0.47	0.596	191.6
SIRST	0.606	0.731	0.951	70.99	0.744	0.841	0.963	275.9
Mixed	0.856	0.92	0.988	1.496	0.758	0.859	0.991	57.66

Since our IRST640 is collected by an IRST system at a fixed location, the images are of a single scene. As the result, if we use only the training set of IRST640, even though the detection performance is excellent on its own test set, it performs poorly on the SIRST test set. This suggests that the network has a tendency of overfitting. When trained with the SIRST, the network has better generalization capability. However, the segmentation performance still needs to be improved. After the mixing, compare with the IRST640 only, a significant increase is achieved on SIRST, although there is a slight decrease in performance on its own test set. Compare with the SIRST only, benefiting from the increasing number of training images, the performance on both test sets gains noticeable improvement. It indicates that the mixed training set can endow the network with stronger generalizability.

6. Conclusions

In this paper, we propose a novel vision transformer-based method for single-frame infrared small target detection, called IRSTFormer. Different from existing methods, this network adopts the pure transformer design for the encoder. Making full use of the self-attention mechanism, the proposed HOPST can learn long-range dependencies in increasingly complex images. For the decoder, a compact module TFAM is presented to perform the feature aggregation progressively. Furthermore, we purpose the CBS loss to supervise the optimization of network parameters after analyzing traditional loss functions in detail. This simple yet effective loss function can bring significant improvement for false alarm suppression. Compared with state-of-the-art methods, our IRSTFormer acquires the best pixel-level and target-level detection performance. The nIoU and detection rate reaches 0.758 and 0.991 on the public SIRST dataset. On our developed IRST640 dataset, our method also has the optimal result. Exhaustive ablation studies demonstrate the effectiveness and reasonability of each component in the method.

In the future, we will focus on the deployment of IRSTFormer in the edge device and explore how to utilize a transformer on multi-frame images for infrared small target recognition.

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

Funding: This research received no external funding.

Data Availability Statement: The IRST640 and SIRST dataset used for training and test are available at: https://github.com/jzchenriver/IRST640 and https://github.com/YimianDai/sirst, accessed on 5 June 2022.

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

References

1. Tartakovsky, A.G.; Kligys, S.; Petrov, A. Adaptive sequential algorithms for detecting targets in a heavy IR clutter. In Proceedings of the Signal and Data Processing of Small Targets 1999, Denver, CO, USA, 4 October 1999; Volume 3809, pp. 119–130.
2. Gao, J.; Guo, Y.; Lin, Z.; An, W.; Li, J. Robust infrared small target detection using multiscale gray and variance difference measures. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 5039–5052. [CrossRef]
3. 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]
4. Tom, V.T.; Peli, T.; Leung, M.; Bondaryk, J.E. Morphology-based algorithm for point target detection in infrared backgrounds. In Proceedings of the Signal and Data Processing of Small Targets, Orlando, FL, USA, 12–14 April 1993; Volume 1954, pp. 2–11.
5. Deshpande, S.D.; Er, M.H.; 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, Denver, Colorado, 20–22 July 1999; Volume 3809, pp. 74–83.
6. Han, J.; Moradi, S.; Faramarzi, I.; Liu, C.; Zhang, H.; Zhao, Q. A local contrast method for infrared small-target detection utilizing a tri-layer window. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1822–1826. [CrossRef]
7. 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]
8. Aghaziyarati, S.; Moradi, S.; Talebi, H. Small infrared target detection using absolute average difference weighted by cumulative directional derivatives. *Infrared Phys. Technol.* **2019**, *101*, 78–87. [CrossRef]
9. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
10. Gao, C.; Zhang, T.; Li, Q. Small infrared target detection using sparse ring representation. *IEEE Aerosp. Electron. Syst. Mag.* **2012**, *27*, 21–30.
11. Dai, Y.; Wu, Y.; Song, Y. Infrared small target and background separation via column-wise weighted robust principal component analysis. *Infrared Phys. Technol.* **2016**, *77*, 421–430. [CrossRef]
12. Wang, H.; Zhou, L.; Wang, L. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8509–8518.
13. 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.
14. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
15. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 593–602.
16. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
19. Ju, M.; Luo, J.; Liu, G.; Luo, H. ISTDet: An efficient end-to-end neural network for infrared small target detection. *Infrared Phys. Technol.* **2021**, *114*, 103659. [CrossRef]
20. Du, J.; Lu, H.; Zhang, L.; Hu, M.; Chen, S.; Deng, Y.; Shen, X.; Zhang, Y. A Spatial-Temporal Feature-Based Detection Framework for Infrared Dim Small Target. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 3000412. [CrossRef]

21. Ding, L.; Xu, X.; Cao, Y.; Zhai, G.; Yang, F.; Qian, L. Detection and tracking of infrared small target by jointly using SSD and pipeline filter. *Digit. Signal Process.* **2021**, *110*, 102949. [CrossRef]
22. Chen, G.; Wang, W. Target recognition in infrared circumferential scanning system via deep convolutional neural networks. *Sensors* **2020**, *20*, 1922. [CrossRef]
23. Fang, H.; Xia, M.; Zhou, G.; Chang, Y.; Yan, L. Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
24. 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.
25. Zhao, B.; Wang, C.; Fu, Q.; Han, Z. A novel pattern for infrared small target detection with generative adversarial network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4481–4492. [CrossRef]
26. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional local contrast networks for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [CrossRef]
27. Hou, Q.; Wang, Z.; Tan, F.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust infrared small target detection network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
28. Zhang, T.; Cao, S.; Pu, T.; Peng, Z. AGPCNet: Attention-Guided Pyramid Context Networks for Infrared Small Target Detection. *arXiv* **2021**, arXiv:2111.03580.
29. 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]
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
31. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
32. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [CrossRef]
33. 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, 18–23 June 2018; pp. 7132–7141.
34. 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.
35. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
36. Dai, Y.; Oehmcke, S.; Gieseke, F.; Wu, Y.; Barnard, K. Attention as activation. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9156–9163.
37. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 3560–3569.
38. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
39. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
40. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning texture transformer network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5791–5800.
41. Zhou, H.Y.; Guo, J.; Zhang, Y.; Yu, L.; Wang, L.; Yu, Y. nnFormer: Interleaved Transformer for Volumetric Segmentation. *arXiv* **2021**, arXiv:2109.03201.
42. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
43. Liu, F.; Gao, C.; Chen, F.; Meng, D.; Zuo, W.; Gao, X. Infrared Small-Dim Target Detection with Transformer under Complex Backgrounds. *arXiv* **2021**, arXiv:2109.14379.
44. Lin, T.Y.; 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.
45. Rahman, M.A.; Wang, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In Proceedings of the International Symposium on Visual Computing, Las Vegas, NV, USA, 12–14 December 2016; pp. 234–244.
46. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense nested attention network for infrared small target detection. *arXiv* **2021**, arXiv:2106.00487.
47. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 821–830.