Toward Accurate Infrared Small Target Detection via Edge-Aware Gated Transformer

Yiming Zhu, Yong Ma, Fan Fan, *Member, IEEE*, Jun Huang, Kangle Wu, and Ge Wang

***Abstract*—Extracting small targets from complex backgrounds is the eventual goal of single-frame infrared small target detection, which has many potential applications in defense security and marine rescue. Recently, methods utilizing deep learning have shown their superiority over traditional theoretical approaches. However, they do not consider both the global semantics and specific shape information, thereby limiting their performance. To overcome this proplem, we propose a gated-shaped TransUnet (GSTUnet), designed to fully utilize shape information while detecting small target detection. Specifically, we have proposed a multiscale encoder branch to extract global features of small targets at different scales. Then, the extracted global features are passed through a gated-shaped stream branch that focuses on the shape information of small targets through gate convolutions. Finally, we fuse their features to obtain the final result. Our GSTUnet learns both global and shape information through the aforementioned two branches, establishing global relationships between different feature scales. The GSTUnet demonstrates excellent evaluation metrics on various datasets, outperforming current state-of-the-art methods.**

***Index Terms*—Gated-shaped stream, infrared small target, Swin Transformer.**

I. Introduction

SINGLE-FRAME infrared small target detection (SISTD) is a critical task that separates small and dim targets from complex backgrounds such as the sky, ocean, and urban structures. It plays an essential role in various fields, encompassing defense security [1], [2], maritime surveillance [2], [3], [4], and precision guidance [2], [5]. Nevertheless, it poses particular challenges. As shown in the red boxes in Fig. 1(a) and (b), small targets occupy only a small portion of the pixels, and their low signal-to-clutter ratios (SCRs) cause them to be susceptible to blending with complex backgrounds [6]. In addition, as shown in Fig. 1(c) and (d), small targets lack texture information, rending traditional object detection methods that focus on this information become not well feasible [7]. Meanwhile, as shown in Fig. 1, the shape and size of targets vary tremendously (5–50 pixels) in different scenarios and backgrounds, potentially resulting in missing detection, false alarm (FA), inaccurate localization, etc. Consequently, SISTD poses a challenging problem. In addressing these challenges, it is necessary to devise a method, particularly on learning shape feature to facilitate the accurate detection of infrared small targets.

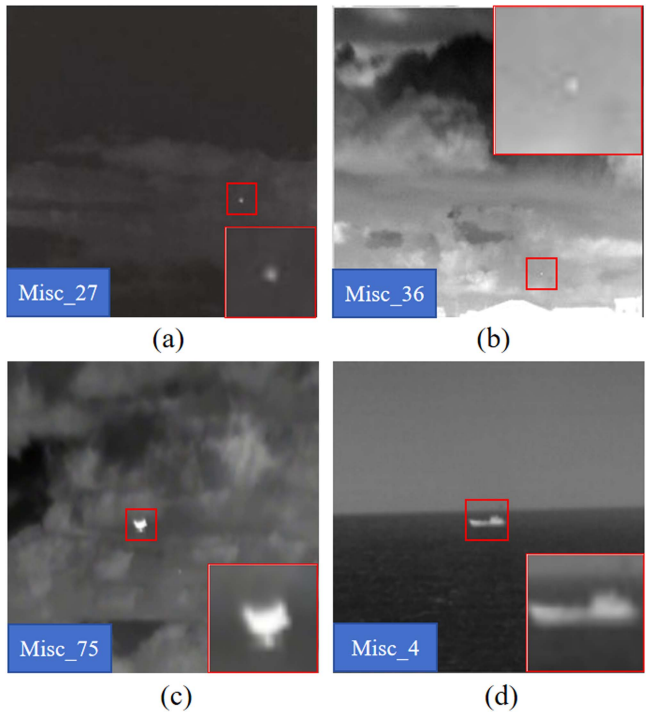

Fig. 1. Unique challenges in SISTD are depicted with the image name from the NUAA-SIRST [8] dataset in the bottom left corner, and the red box represents a local zoom. (a) Small and gray point target against terrestrial background. (b) Point target with low SCRs against sky background. (c) Small target without texture information but with distinct shape in sky background. (d) Small targets with discernible shape information against ocean background.

Traditional SISTD methods can be classified into three classes, including filter-based, human vision system (HVS)-based, and low-rank matrix (LRM)-based methods. Filter-based methods [9], [10], [11], [12] employ specifically designed filters to extract small target from the backgrounds. While they are effective in filtering out smooth background clutter, their performance degrades significantly when encountering noise and background interference of varying intensity. HVS-based methods [6], [7], [13], [14], [15] rely on the local brightness and darkness contrast difference between targets and backgrounds, making them particularly suitable for detecting small targets

Manuscript received 29 January 2024; revised 27 March 2024; accepted 7 April 2024. Date of publication 10 April 2024; date of current version 29 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62075169, Grant 62003247, and Grant 62061160370, and in part by the Key Research and Development Program under Grant 2021BBA235 of Hubei Province. *(Corresponding author: Fan Fan.)*

The authors are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: yiming_zhu_2015@163.com; mayong@whu.edu.cn; fanfan@whu.edu.cn; junhwong@whu.edu.cn; wukangle80@gmail.com; wangge_phd@whu.edu.cn).

Data is available online at: https://github.com/yimingzhu2015/GSTUnet_infrared_det

Digital Object Identifier 10.1109/JSTARS.2024.3386899

with relatively high brightness. However, in the presence of strong bright noise in the background, they may result in higher FA rates. LRM-based methods [16], [17], [18], [19], [20] treat the infrared image as a low-rank sparse matrix, and introduce the LRM reconstruction to filter the targets with the backgrounds. Since the intensity of the target is not significant with respect to the intensity of the backgrounds, these methods do not detect targets under various shapes well. Overall, when dealing with complex scenarios, the traditional methods heavily rely on manual features, which makes it difficult to cope with complex scenes.

Deep learning methods [21], [22], [23], [24], [25], [26], [27] have revolutionized the field of SISTD by applying data-driven strategies, which are widely used due to their ability to learn infrared small target features. Among deep learning-based methods, both CNN-based and Transformer-based approaches have been explored. CNN-based methods have achieved excellent results by improved convolution for small target scenarios. However, the convolution operations have limited local receptive fields [28], and the introduced pooling operations will ignore small targets during the downsampling process [24]. Transformer-based approaches address these issues by integrating global receptive fields and self-attention mechanisms, but they currently do not specifically consider local feature information such as edge and shape, which may not only lead to miss detection, but also affect the recognition of target types in practical applications [e.g. Fig. 1(c) and (d) can be further identified as UAV and ship]. Due to the low contrast and low SCR between infrared small targets and the background, it is challenging to extract useful edge and shape features of the targets. Thus, obtaining accurate edges and shapes of infrared small targets remains a challenging task.

To overcome the aforementioned drawbacks, we consider to explore a novel paradigm of incorporating target shape learning into SISTD. we propose gated-shaped TransUnet (GSTUnet) for SISTD through semantic segmentation. The feature extractor of GSTUnet consists of two key branches, in the first branch, we proposed a multiscale feature extraction network built on Swin Transformer [29] to overcome the receptive field limitation of convolution operations. This enables the extraction of global semantic features of small targets at different scales. In the second branch, we design an edge-aware gated-shaped stream that focuses specifically on capturing the shape and edge information of small targets by using a gated convolution layer composed of two convolution layers and a sigmoid layer with ResNet [30]. Furthermore, GSTUnet integrates the semantic and shape information from the aforementioned two branches and establishes global dependencies among different feature scales, allowing for highly discriminate detection of small infrared targets, and to obtain the exact shape of the target. The contributions of our work can be summarized as follows.

1) We propose a novel architectural paradigm that incorporates and meticulously considers edge and shape information within the ambit of SISTD.
2) We have engineered an edge-aware gating mechanism architecture that adeptly extracts edge and shape features of diminutive targets.
3) We propose an edge-aware loss function specifically tailored to accentuate the extraction of edge and shape information against complex backgrounds while concurrently suppressing FAs.
4) We have conducted extensive experiments on relevant benchmarks, substantiating the efficacy of our method, which has culminated in state-of-the-art (SOTA) results.

II. Related Work

A. Single Infrared Small Target Detection

Traditional SISTD methods are based on maintaining consistency in backgrounds and enhancing the salience of small infrared targets, relying mainly on manual features, suFch as HVS-based, filter-based, and LRM-based methods.

HVS-based methods primarily rely on the salience of local gray value contrast between small targets and backgrounds. For instance, Chen et al. [13] introduced the local contrast measure (LCM), and Wei et al. [14] further considered the multiscale feature extraction and designed a multiscale patches contrast module (MPCM) to enhance the LCM for computing the salience of small targets salience. However, MPCM neglects global information, prompting Qiu et al. [6] to develop the global structure-based LCM (GSLCM), which appropriately incorporates global information. Building upon GSLCM, Qiu et al. [7] also proposed an adaptive structure patch contrast measure (ASPCM). However, the performance of HVS-based methods degrades significantly in the presence of noise or clutter in the image.

Filter-based methods focus on suppressing background clutter, For instance, Bai and Zhou [9] proposed the top-hat filter as a filtering-based method. Li et al. [12] utilized both max-mean/max-median filter to detect small targets. However, this type of methods have some limitations, such as the model parameters enable automatic adjustment on complex backgrounds, making it suitable only for single backgrounds and uniform scenes.

LRM-based methods have been developed to utilize the properties of targets and backgrounds for decomposing the original image into target and background components. For instance, Dai and Wu [16] proposed the reweighted infrared patch tensor, and Wu et al. [31] proposed a tensor train/ring expansion methods, which utilize alternating direction multiplier method to expand the components by weighting the kernel norm and balancing the constraints. While LRM-based methods can achieve superior detection performance, they often require multiple iterations to converge and attain optimal solutions, resulting in a less efficient process.

Traditional methods are highly interpretable and can provide satisfactory results without the need for large amounts of data. However, they rely on prior assumptions and exhibit low robustness in complex backgrounds.

B. CNN-Based Infrared Detection Framework

With the development of CNN, which enable the extraction of features from data-driven and provide end-to-end processing,

CNN-based methods have surpassed traditional approaches in suppressing FAs, more precision, and enhancing robustness. For instance, Ju et al. [32] used ISTDet, which consists of an image filtering module and an infrared small target detection module. Dai et al. [8] proposed a one-stage cascade refinement network that cascades the CNN features for detecting small targets in infrared images. Qian et al. [33] proposed SiamIST, which combines side window filter with the improved SiamRPN model. Han et al. [3], [4] proposed the efficient information reuse network in marine backgrounds, which combines a dense feature fusion network with two fusion directions in feature extraction, and a dual mask attention module to refine the fused feature map. Meanwhile, facing the special situation of small samples, they proposed a balanced feature fusion network and a context attention network. However, object detection techniques only provide bounding box information and lack detailed shape information, which is important for further analysis and understanding of target characteristics in the SISTD.

Since the semantic segmentation can provide detailed segmentation results at the pixel level, where the shape of targets can be more accurately detected, addressing the aforementioned issue. Therefore, some works modeled SISTD as a semantic segmentation task. [23], [24], [25], [27], [34], [35], [36], [37], [38], [39], [40]. For instance, Dai et al. [23], [34] considered the feature of small targets and constructed upsampling asymmetric contextual modulation (ACM) and attention local contrast (ALC) modules, integrating them into CNN structure. Zhang et al. [38] proposed attention guided pyramid context network, utilizing attention mechanisms to explore contextual information and preserve detailed information. Kou et al. [39], [40] proposed a multistrategy fusion model that integrates various class convolution modules, including postprocessing of eight neighborhood clustering, achieved real-time infrared small target detection and tracking. As for the shape feature learning, Zhang et al. [24] designed an end-to-end Taylor differential operator for improved edge and shape detection. Lin et al. [41] proposed a framework to fully consider shaped-biased learning for accurate detect infrared small target. However, the aforementioned methods are all based on CNN, which suffers from a limited ability to focus on global information due to convolution and multiple downsampling operations through pooling in the network. To address this issue, a suitable feature extractor of framework should be employed.

C. Transformer-Based Infrared Detection Framework

The vision transformer (ViT) [42] has proven highly effective in computer vision tasks. To extract cross-windows features, the Swin Transformer [29] used sliding window attention and has achieved satisfactory results such as image classification [43], [44], object detection [26], [45], [46], and semantic segmentation [47], [48]. The Swin Transformer consists of two series blocks [29]. The first block, shown in the left part of Fig. 2, utilizes a W - MSA module, whereas the second block, shown in the right part of Fig. 2, utilizes the SW - MSA module. The W - MSA and SW - MSA modules employ windows attention and sliding windows attention, respectively. The W - MSA layer

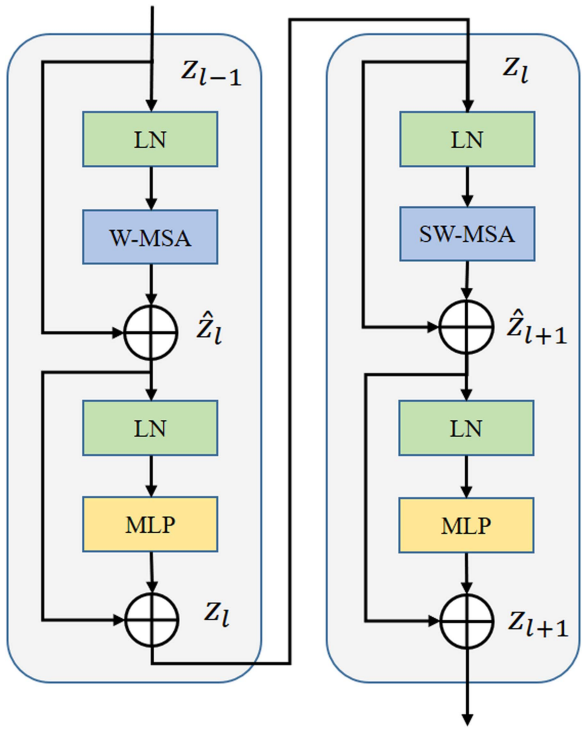

Fig. 2. Two successive Swin Transformer blocks. W-MSA and SW-MSA are the multihead self-attention modules with windows and sliding windows self-attention.

separates the input infrared image into nonoverlapping windows, with each window having a size of $M \times M$. The different windows splitting strategies aim to minimize the effects of windows localization, as illustrated in Fig. 4. The W - MSA module utilizes localized windows to compute the self-attention of each individual window, as shown in the left part of Fig. 4. The SW - MSA layer is designed to facilitate the creation and extraction of adequate global information between windows without introducing additional computation costs. The SW - MSA layer employs the sliding window method, as shown in the right part of Fig. 4. The SW - MSA employs the sliding window method and iterates from upper left to bottom right. In addition, the Swin Transformer module consists of a multilayer perception (MLP) module, a layer-norm (LN) module, and residual connections. The computational procedure for two consecutive Swin Transformers is as follows:

$$\begin{aligned} \hat{z}_{l+1} &= \text{SW - MSA}(\text{LN}(z_l)) + z_l \\ z_{l+1} &= \text{MLP}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1} \\ \hat{z}_l &= \text{W - MSA}(\text{LN}(z_{l-1})) + z_{l-1} \\ z_l &= \text{MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l \end{aligned} \tag{1}$$

where $\hat{z}_l$ represents the output of W - MSA or SW - MSA module of lth block and z_l represents the output of MLP module.

The Swin Transformer computes self-attention from the feature map using batch windows. Each batch is composed of multiple nonadjacent subwindows, as shown in Fig. 4. The patches that include small targets are marked with red bounding boxes,

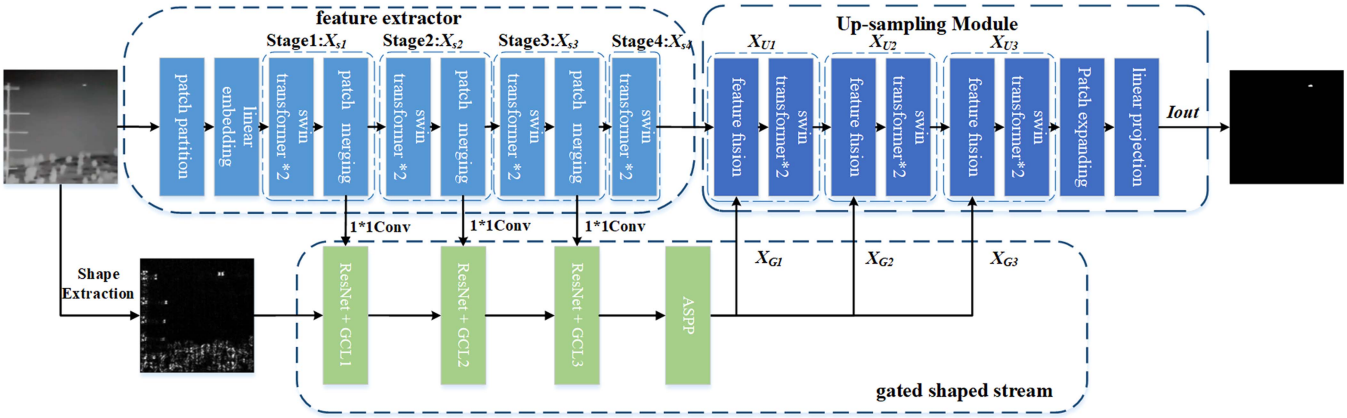

Fig. 3. Overall structure of our GSTUnet, the upper left is a multiscale feature extraction module, the bottom part is a gated-shaped stream, and the upper right is feature fusion module and upsampling module.

and self-attention can be computed in different subwindows. The calculation of self-attention in a subwindow is as follows:

$$\mathbf{Q},\mathbf{K},\mathbf{V}=\{\mathbf{T}\mathbf{W}^{q},\mathbf{T}\mathbf{W}^{k},\mathbf{T}\mathbf{W}^{v}\}$$
$$\text{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_{T}}{\sqrt{d}}+\mathbf{B}\right)\mathbf{V} \quad (2)$$

where $\mathbf{T}\in\mathbb{R}^{M^2\times d}$ represents the input of self-attention, and $\mathbf{W}^{q},\mathbf{W}^{k}$, and $\mathbf{W}^{v}\in\mathbb{R}^{d\times d}$ represent the weights of three linear projection layers, which implemented using an MLP, $\mathbf{Q},\mathbf{K}$, and $\mathbf{V}\in\mathbb{R}^{M^2\times D}$ indicates the query, key, and value matrices, respectively. M^2 denotes the size of $\mathbf{T}$, $\mathbf{Q}$, and $\mathbf{K}$, and $\mathbf{V}$ and $\mathbf{D}$ denote the channel dimensions. In Swin Transformer, a bias matrix $\hat{\mathbf{B}}\in\mathbb{R}^{(2M-1)\times(2M-1)}$ is parameterized, and the values of $\mathbf{B}$ are taken from this bias matrix.

Building upon ViT for SISTD, Lin et al. [26] proposed U-Transformer based on Swin Transformer and CNNs, which has demonstrated satisfactory results in SISTD. Lin et al. [49] proposed IST-TransNet which combines ViT with CNN blocks antialiasing contextual feature fusion module and spatial and channel attention module. Liu et al. [50] proposed a ViT-based feature embedding module with a feature enhancement module to learn discriminate features of small targets and prevent missed detection. These findings suggest that combining CNN with Transformer models holds great potential for feature extraction and can address the limitations of CNNs. However, the above transformer-based methods do not specially consider the local shape information of small targets. Drawing inspiration from the above methods, our approach has explored a new idea of incorporating target shape reconstruction into small infrared target detection, which significantly improves over existing methods and holds great potential for advancing the field of SISTD.

III. Methodology

The overall structure is illustrated in Fig. 3. GSTUnet comprises of three components: a multiscale feature extraction module, a gated-shaped stream module, and a feature fusion module. In the multiscale feature extraction module, we utilize four sequential stages to extract multiscale semantic feature maps (see Section III-A). The gated-shaped stream is designed to extract the shape information from small targets through gate convolutions (see Section III-B). In the feature fusion module, we fuse the semantic feature with shape features at multiscale to obtain the final results (Section III-C). In Section III-D, we present the loss function for end-to-end training of our network.

The proposed GSTUnet aims to generate detected binary mask $\mathbf{I}_{\text{out}}\in\mathbb{R}^{H\times W\times 1}$ to indicate the detected regions. Specifically, let $\mathbf{I}_{\text{in}}\in\mathbb{R}^{H\times W\times C_{in}}$ represent the dimensions of the input infrared image. Here, H, W, and C_{in} represent the height, width, and channels of the input image, respectively.

The notation used in this article is as follows: $\boldsymbol{X}_{\boldsymbol{S}_1},\boldsymbol{X}_{\boldsymbol{S}_2},\boldsymbol{X}_{\boldsymbol{S}_3}$, and $\boldsymbol{X}_{\boldsymbol{S}_4}$ represent the feature maps of each encoder stage. The feature in the Swin Transformer is denoted as z_l, whereas the $\boldsymbol{X}_{\boldsymbol{G}_i}$ represents the different size feature map of the gated-shaped stream, the upsampling feature maps are $\boldsymbol{X}_{\boldsymbol{U}_1},\boldsymbol{X}_{\boldsymbol{U}_2},\boldsymbol{X}_{\boldsymbol{U}_3}$, and $\boldsymbol{X}_{\boldsymbol{U}_4}$.

A. Multiscale Feature Extraction Module

The detection of infrared small targets in the presence of complex and variable background becomes increasingly challenging due to the scale variation. Therefore, feature extraction must consider this issue. And since multiscale feature representation has been shown to be effective in adapting the scale variation [6], [8], [51], we incorporate a multiscale mechanism in the design of our feature extractor. Moreover, Swin Transformer preforms well in SISTD due to its excellent global attention mechanism and semantic feature extraction capability. Therefore, we design the feature extractor using multiple Swin Transformers, which are capable of extracting multiscale global semantic features.

The feature extractor consists of four stages. Each stage contains two Swin Transformers blocks, that utilize W-MSA and SW-MSA to capture self-attention for windows and sliding windows, respectively. This allows for a global receptive field view during self-attention computation, and more details have been illustrated in Section II-C. Subsequently, the output feature maps of the Swin Transformer are downsampled at each stage

to capture multiscale global semantic information. The general approach for downsampling is through maximum or average pooling. However, this may result in the loss of detailed features for small targets when the targets are too small [25]. Therefore, we utilized patch merging instead of pooling to downsample the feature map. Patch merging has better performance in feature retention as it integrates all the feature information of each image patches.

It is worth noting that the self-attention calculations in the Swin Transformer, as shown in (2), are complicated. Feeding the input image directly to the multiple Swin Transformer is computationally complex and time-consuming. Therefore, we sequentially introduce a patch partition (PP) layer and a linear embedding (LE) layer before it. The PP layer transforms the input image $\mathbf{I}_{\text{in}} \in \mathbb{R}^{H\times W\times C_{\text{in}}}$ into image patches to reduce computational complexity. Then, the LE layer map the channels of each patch to a specified dimension C. Specifically, the resolution of each patch is $H/4 \times W/4$, and convolution is utilized in the PP layer and the LE layer. The convolution kernel size and stride are both $H/4$, the input channels of the infrared image are 1, and the output channels of the LE layer channel are C.

Overall, the multiscale feature extraction module aims to capture detailed information and adapts to different target scales to enhance the performance of SISTD. The structure is shown in the top-left of Fig. 3. First, it utilizes PP and LE layer to transform the input image into patches and obtain an initial feature map. This step can be expressed as follows:

$$\boldsymbol{X}_{\boldsymbol{S}_0} = \text{Conv}_{\text{PPandLE}}(\mathbf{I}_{\text{in}}) \tag{3}$$

Second, the initial feature map is fed into the four-stage Swin Transformer to calculate the self-attention of each window, and then downsampling the feature map using a patch merging layer. The step can be expressed as follows:

$$\boldsymbol{X}_{\boldsymbol{S}_i} = \text{SwinStage}[\boldsymbol{X}_{\boldsymbol{S}_{i-1}}](i = 1, 2, 3, 4) \tag{4}$$

where $\boldsymbol{X}_{\boldsymbol{S}_i}(i = 1, 2, 3, 4)$ is the output feature maps of each of Swin Transformer stages. At each stage, the output feature map size is halved compared with the input feature map size, whereas the channel size is doubled. After the four stages, the size of the output feature map decreases from $(H/4) \times (W/4)$ to $(H/32) \times (W/32)$, and the channel size increases from C to $8\,C$.

Finally, these four feature maps are concatenated along the channel dimension to produce a feature map of size $(H/2^{i+1}) \times (W/2^{i+1}) \times 2^{i+1}C$, which is then transformed to 2^iC channels using a linear projection layer.

Specifically, at the ith stage, interval sampling is performed on the input feature map $X_{S_i} \in \mathbb{R}^{(H/2^{i+1})\times(W/2^{i+1})\times 2^{i-1}C}$, resulting in four feature maps of size $(H/2^{i+1}) \times (W/2^{i+1}) \times 2^{i-1}C$. The output feature maps of each stage are referred to as $\boldsymbol{X}_{\boldsymbol{S}_1}, \boldsymbol{X}_{\boldsymbol{S}_2}, \boldsymbol{X}_{\boldsymbol{S}_3}$, and $\boldsymbol{X}_{\boldsymbol{S}_4}$.

B. Edge-Aware Gated-shaped Stream Module

The environment backgrounds are dynamic and perplex, with various types and orientations of targets, leading to diverse shapes of small targets [41]. The previously extracted multiscale features are obtained by processing the semantic features of the context in the patch sequence. However, they lack the ability to capture the shape of the target since it does not directly process the spatial information. This results in the direct extraction of small targets from the multiscale features, which leads to a network that lacks both robustness and sensitivity to changes in target shape. Therefore, extracting shape information can increase the robustness of SISTD.

For this reason, we design a second auxiliary branch, called the edge-aware gated-shaped stream, aiming to accurately extract the shape features of small targets and improving the detection accuracy. It is an end-to-end network and its structure is illustrated in Fig. 5. It comprises gated convolutional layer (GCL), ResNet [30], and 1×1 convolution and ASPP [52] modules. The details of this structure are as follows.

First, considering that the edges are important structural elements in images that implies the shape information of the target, it is essential to extract the shape features from the edge region. Therefore, we utilize the Sobel operator [53] to extract edges from the original infrared image, resulting in the edge prior graph $\boldsymbol{X}_{\boldsymbol{E}}$. Next, $\boldsymbol{X}_{\boldsymbol{E}}$ is fed into the gated-shaped stream through the input arrow in the lower left corner of Fig. 5.

$$\boldsymbol{X}_{\boldsymbol{E}} = \text{Sobel}(\boldsymbol{I}) \tag{5}$$

Then, previous studies [24], [41] have demonstrated that incorporating contextual semantic information enhances the ability of model to perceive edges. Based on this insight, we feed multiscale feature maps, comprising $\boldsymbol{X}_{\boldsymbol{S}_1}$, $\boldsymbol{X}_{\boldsymbol{S}_2}$, and $\boldsymbol{X}_{\boldsymbol{S}_3}$, into the gated-shaped stream, facilitating the fusion of the edge information and the semantic feature information. Moreover, to extract the local details and edge information of small targets and improve the target detection performance, we further extract deep features from these multiscale feature maps. In this process, considering that the residual network can prevent small target features from being lost during downsampling [25], we concatenate 1×1 convolution and ResNet [30] after $\boldsymbol{X}_{\boldsymbol{S}_i}$. The network extracts deep features of multiscale feature maps $\boldsymbol{X}_{\boldsymbol{S}_1}$, $\boldsymbol{X}_{\boldsymbol{S}_2}$, and $\boldsymbol{X}_{\boldsymbol{S}_3}$ by concatenating the residuals, respectively. This deep feature extraction formula is as follows:

$$\boldsymbol{X}'_{\boldsymbol{S}_i} = \text{ResNet}(\text{Conv}_{1*1}(\boldsymbol{X}_{\boldsymbol{S}_i})), (i = 1, 2, 3) \tag{6}$$

where ResNet denotes the residual network, $\boldsymbol{X}_{\boldsymbol{S}'_i}$ denotes the feature after deep feature extraction. As a deep ResNet may result in extensive computational requirements, we employ a lightweight ResNet18 module. Its parameters are obtained from the pretrained model weights on ImageNet [54], and the three ResNet18 networks share the same set of weights.

After acquiring the deep semantic features $\boldsymbol{X}'_{\boldsymbol{S}_i}$ and the edge map $\boldsymbol{X}_{\boldsymbol{E}}$, the next step is to fuse them. It is important to note that the edge information of infrared small targets in semantic features tends to be blurred [24]. Consequently, directly fusing the edge maps with deeper features can introduce noise and redundancy. To address this, we design the GCL module to selectively process edge-related information in deep semantic features and fuse them level by level. The input of the first level GCL consists of $\boldsymbol{X}_{\boldsymbol{E}}$ and $\boldsymbol{X}_{\boldsymbol{S}_1}$. The inputs to the second and third

GCLs are composed of the outputs of the previous GCLs and the corresponding multiscale deep semantic features, respectively.

The structure of GCL is illustrated in Fig. 6, which first concatenates the deep semantic features and edge features. As shown in the blue box on the left side of Fig. 6, where the edge semantic features (represented by light green blocks) and deep semantic features (represented by dark green blocks) are combined. Next, the stacked features are processed through two consecutive 1×1 convolution and batch normalization operations, both followed by ReLU activation function. Finally, the sigmoid function to implement the gated convolution operation. In addition, we integrate the two residual connections into the input feature map by elementwise multiplication and elementwise addition, the formula for GCL is as follows:

$$\begin{aligned} \boldsymbol{\alpha_i} &= \sigma(\mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{1*1}(\boldsymbol{X_i}||\boldsymbol{Y_i})))) \\ \boldsymbol{X}_{\boldsymbol{S_i}}^{\mathrm{out}} &= \mathrm{Conv}_{1*1}(\boldsymbol{X_i} \odot \alpha_i + \boldsymbol{X_i}), (i = 1, 2, 3) \end{aligned} \tag{7}$$

where $\boldsymbol{X_1}$ denotes the edge prior graph $\boldsymbol{X_E}$, $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$ denote the outputs from the previous level of GCL, Y_i denotes the feature graph after deep semantic feature extraction, BN represents batch normalization, ReLU denotes the activation function, σ denotes the sigmoid function, $\odot$ denotes elementwise multiplication, $||$ denotes the concating of the feature map, and $\boldsymbol{X}_{\boldsymbol{S_i}}^{\mathrm{out}}$ denotes the output of the GCL module. When in the flat nonedge region, the value of α_i will be relatively small, and the formula will be calculated to get relatively small $\boldsymbol{X}_{\boldsymbol{S_i}}^{\mathrm{out}}$, making the corresponding output of the gated-shaped stream be relatively small. In summary, GCL controls the flow of information by using sigmoid function as a gating unit, which can effectively filter out the details of flat regions in the feature map and put more focus on image features such as edges and shapes, and therefore, can output feature maps that contain only edge information.

Finally, to achieve multiscale feature fusion during the upsampling process of the extracted feature maps contain edge information, we generate multiscale edge feature maps by upsampling the feature maps containing edge features with the purpose of fusing edges and small targets. Specifically, we introduce the ASPP [52] module to upsample the output of the final level of GCL to match size as the $\boldsymbol{X_{S_1}}$, $\boldsymbol{X_{S_2}}$, and $\boldsymbol{X_{S_3}}$, as shown in Fig. 7. In addition, a 1×1 convolution is applied to match the output channel with the input channel. We represent the feature map after upsampling, as $\boldsymbol{X_{G_1}}$, $\boldsymbol{X_{G_2}}$, and $\boldsymbol{X_{G_3}}$.

C. Upsampling Module

To integrate multiscale features and shape information of the target, our upsampling process is illustrated in the upper right corner of Fig. 3. The upsampling process comprises three stages, each including a fusion module and two Swin Transformer modules. First, we introduce a fusion module that integrates features from multiscale feature extraction and the output characteristics of the gated-shaped stream. The fusion block employs a fully connected layer and a linear projection layer to merge and upsample the features, thereby, reducing the feature dimensional. Second, the Swin Transformer module is designed to learn finer feature relationships effectively.

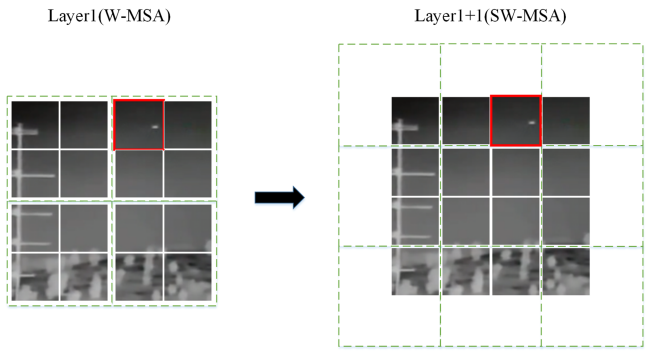

Fig. 4. Window partition strategy of W-MSA module and SW-MSA. (Red rectangle is the patch which includes small target.)

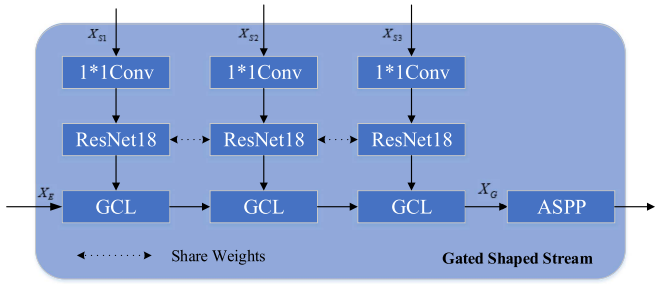

Fig. 5. Specifically of gated-shaped stream, which includes 1×1 convolution, ResNet, GCL, and ASPP.

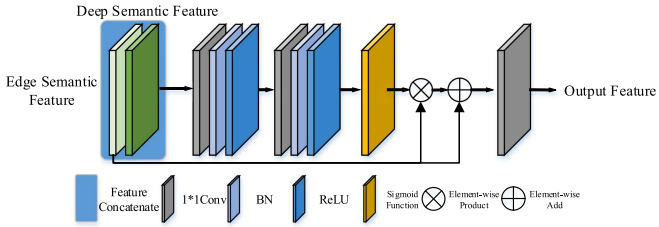

Fig. 6. Specifically of gate convolution layer, input edge semantic feature and deep semantic feature, with 1×1 Conv, BN, and ReLU.

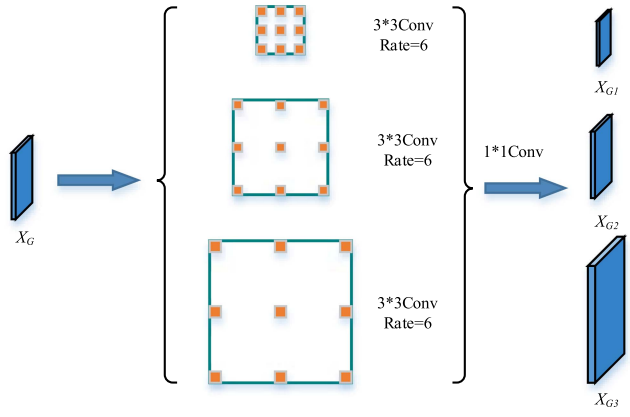

Fig. 7. Specifically of ASPP.

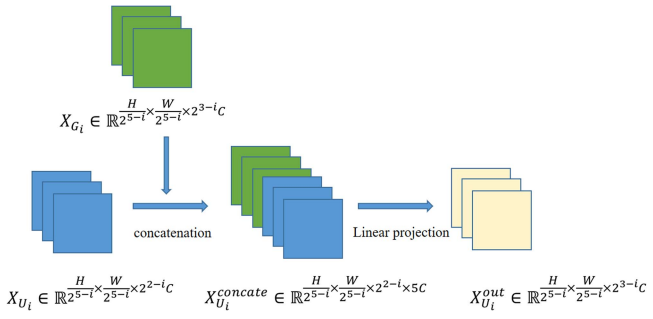

Fig. 8. Structure of an upsampling stage.

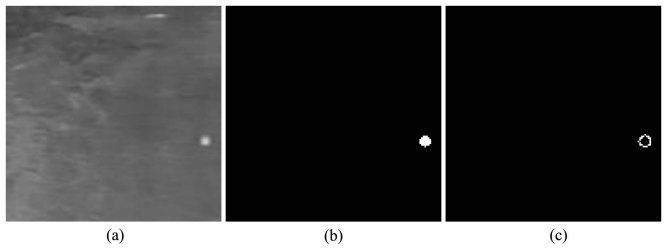

Fig. 9. (a) Input infrared image. (b) Ground truth of feature extractor. (c) Gated-shaped stream.

The feature maps of each stage are denoted as $\boldsymbol{X}_{\boldsymbol{U}_i}$ (where $i = 1, 2, 3$). To mitigate information loss caused by upsampling, the feature fusion module comprises a fully connected MLP and a linear projection F_{proj}, as shown in the Fig. 8. The fusion block utilizes a patch reshape block to perform the upsampling operation. Subsequently, the upsampled feature maps $\boldsymbol{X}_{\boldsymbol{U}_i}$ (where $i = 1, 2, 3$) are concatenated with the feature maps $\boldsymbol{X}_{\boldsymbol{G}_i}$ generated by the gated-shaped stream. Afterward, F_{proj} is applied to increase the dimensionality of $\boldsymbol{X}_{\boldsymbol{U}_i}$. The dimension is increased from $2^{4-i}C$ to $2^{5-i}C$, where C represents the channel dimension. The upsampling process is formulated as follows:

$$X_{Ui+1} = F_{\text{proj}}[\text{MLP}[\text{Concat}(X_{Ui}, X_{Gi})]]. \tag{8}$$

During the upsampling process, after each stage, the channel dimension of $\boldsymbol{X}_{\boldsymbol{U}_i}$ is halved compared with $\boldsymbol{X}_{\boldsymbol{U}_{i-1}}$, and the spatial resolution is doubled. The size of the feature map increases from $H/32 \times W/32$ to $H/4 \times W/4$, whereas the channels decreases from $8\,C$ to C. The ultimate output is the detection image I_{out}, where $I_{\text{out}} \in \mathbb{R}^{H \times W \times 1}$.

D. Edge-Aware Loss Function Design

GSTUnet is trained end-to-end with a hybrid edge-aware loss function. During the training, we proposed a loss function to supervise the output and the boundary map. Here, we utilize the Sobel operator on the ground truth to indicate the contours of the small targets. We follow the standard binary cross-entropy (BCE) loss [24] on the predicted boundary map s, and the BCE loss on the predicted semantic output map s. Our loss function comprises two components: one for the output and the other for the gated-shaped stream, which ground truth is obtained by applying the Soble [55] operator to the ground truth of the input image, as shown in Fig. 9. We simultaneously train two data streams, incorporating both the loss function for semantic segmentation and the loss function for small target boundaries. The loss function for the output map is a combination of Dice [56] loss and BCE loss functions.

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \lambda_1 \mathcal{L}_{\text{BCE}}(\nabla y, X_S^{\text{out}}) + \mathcal{L}_2(y, X_{\text{out}}) \\ \mathcal{L}_2 &= \lambda_2 \mathcal{L}_{\text{Dice}}(y, X_{\text{out}}) + \lambda_3 \mathcal{L}_{\text{BCE}}(y, X_{\text{out}}) \end{aligned} \tag{9}$$

where the ∇y is the edge ground truth. X_S^{out} is the output of the gated-shaped stream, y is the ground truth, X_{out} is the output map of GSUnet, and X_S^{out} is the output of the gated-shaped stream.

IV. Experimental Results and Analysis

In Section IV, we conducted experiments on two infrared datasets and compared the results of SOTA methods to evaluate the learning and generalization capabilities of GSTUnet. In Section IV-A, we first provide an overview of the datasets. In Section IV-B, we describe implementation details. In Section IV-C, we introduce the evaluation metrics and thoroughly analyze the experimental results both quantitatively and qualitatively. In addition, we provide ablation studies for the effectiveness of each module.

A. Datasets

1) Infrared Datasets Descriptions: To validate the methodology of this article, we evaluate our method on two public datasets: NUAA-SIRST [23] and NUDT-SIRST [25].

NUAA-SIRST is a small open-source SISTD dataset, which contains 427 representative images from hundreds of real-world videos captured at short-midwave-950 nm wavelength, annotated in five ways, such as VOC annotation and mask annotation, to support image detection and semantic segmentation or instance segmentation tasks. Most small infrared targets are dim and hidden in complex backgrounds such as the sky, ocean, and buildings. In addition, nearly 40% of the images belong to the brightest pixels.

NUDT-SIRST is a large open SISTD dataset, which is a synthetic dataset and contains 1327 images created by collecting high-resolution natural scene images, cropping different areas from these images to get different backgrounds, and then overlaying small targets by using a 2-D Gaussian function onto the backgrounds to form synthetic images.

2) Data Augmentation Techniques: We leverage data augmentation to generate additional instances from raw data via transformations, including rotation, translation, and scaling. Our experiment involved rotating the dataset by 90°, 180°, and 270°, and applying random cropping and scaling to diversify the dataset further. Given the significant contrast discrepancies observed in infrared small target detection datasets across various backgrounds, we employed image brightness and contrast perturbation techniques for data augmentation. The data augmentation methods can enhance the quantity of training data and improve the network's generalization capabilities.

TABLE I
TRAINING PARAMETERS

Training Parameters	Details
Framework	PyTorch1.8.1
Operating System	Ubuntu
GPU	RTX3090Ti
Accleration Tools	CUDA11.3
Epochs	200
λ_1	1
λ_2	1
λ_3	0.5
Learning rate	2×10^{-4}
Weight Decay	1×10^{-5}
Optimizer	Adam

More details are shown in GitHub Project.

B. Experimental Details

GSTUnet was implemented by the PyTorch framework and trained on NVIDIA RTX 3090Ti GPU, running on the Ubuntu environment. The model was trained for 200 epochs. Simultaneously, we applied gradient clipping and cosine annealing to assist the network training. The specific training parameters are listed in Table I. Our model has three variants: small-scale, basic-scale, and large-scale versions, and they use Swin-Tiny, Swin-Base, and Swin-Large as feature extraction encoders, respectively. These Swin Transformer models use the pretrained model weights from [29]. Different scales of Swin Transformer blocks, including many layers and attention layers, are shown in Table II.

C. Experiment Results

In Section IV-C, we compare the performance of GSTUnet with other methods. We present quantitative results alongside qualitative visualizations in different backgrounds and target scales to evaluate our method. Furthermore, we conducted an ablation study to analyze the contribution of individual technical components in GSTUnet.

1) Evaluation Metric: To evaluate shape description ability, our experiments utilize intersection over union (IoU) and normalized intersection over union (nIoU). Meanwhile, we adopt probability detection (Pd) and FA to evaluate the localization and detection ability. Furthermore, we draw receiver operation characteristics (ROC) curves to describe how detection rates evolve with varying FA rates.

a) *IoU:* IoU is a pixel-level evaluation metric to describe the contour of a small target, which calculates the ratio of intersection area and union area between the prediction map of small targets and the pixel annotated ground truth, i.e.,

$$\text{IoU} = \frac{A_i}{A_u} \tag{10}$$

where A_i and A_u denote the size of the intersection and union area of the infrared small target prediction mask and ground truth, respectively.

b) *nIoU:* nIoU is the normalized IoU, which is calculated as follows:

$$\text{nIoU} = \frac{1}{N}\sum_{i=1}^{N}(\text{TP}[i]/(T[i] + P[i] - \text{TP}[i])) \tag{11}$$

where N is the total number of datasets, TP denotes the number of true positive pixels, T and P denote the number of ground truth and predict positive pixels, respectively.

c) *Pd:* Pd denotes the probability of detection, and we specified that if the deviation of the predicted mask's regional center and the center of the ground truth is less than a threshold (set to 3 in this article), these targets are considered as correctly predicted targets, which is calculated as follows:

$$P_d = \frac{N_{\text{pred}}}{N_{\text{all}}} \tag{12}$$

where N_{pred} is the correctly predicted point, and all targets number is the N_{all}.

d) *FA:* FA denotes the false alarm rate. We specified that if the deviation of the predicted mask's regional center and the center of the ground truth is more than the predefined threshold (set 3 in this article), we consider those pixels as falsely predicted ones

$$F_a = \frac{N_{\text{false}}}{N_{\text{all}}} \tag{13}$$

where N_{false} is the false alarm points, and all targets number is the N.

e) *ROC:* ROC curve describes the change tendency of the Pd under different FA. The area of the ROC curve expresses the robustness of the model, with a larger area indicating a stronger model.

2) Compare With SOTA Methods: Our comparison include quantitative results and qualitative results.

a) *Quantitative results:*We select MDvsFAcGAN [27], ALCNet [34], ACMNet [23], IAANet [36], UIUNet [35], and CGRANet [37] as the CNN-based methods. To compare the traditional methods, we select the filter-based method SRTHT [11], max-mean/median-based method TLMS [12], the HVS-based method MPCM [14], and the LRM-based methods: IPI [17] and RIPT [16]. The key parameters configuration are shown in Table III:
The quantitative results are shown in Table IV, we selected the top-performing from the three GSTUnet variants (-T, -B, and -L) as our method results. Our quantitative results demonstrate SOTA performance in Table IV. Our method achieves 82.61% and 84.51% in the IoU and nIoU metrics and achieves 80.97% and 82.19% in the nIoU, respectively, indicating the effectiveness of our approach in detecting the shapes of small targets. It also achieves 98.91% and 99.91% in Pd, respectively, compared with the lower FA of 5.32×10^{-6} and 4.92×10^{-6}, reflecting the Pd of small targets with fewer FA enabled by our method.
We also plot ROC and Precision-Recall curves for comparison experiments, as Fig. 10, which demonstrates that GSTUnet significantly outperforms other methods, with a

TABLE II
DETAILS OF SWIN TRANSFORMER BACKBONE VARIANTS

Models	Hidden Size C	MLP Size D	Layer Num	Head Num	Window Size
Swin-Tiny	96	384	[2, 2, 6, 2]	[3, 6, 12, 24]	7
Swin-Base	128	512	[2, 2, 18, 2]	[4, 8, 12, 24]	7
Swin-Large	192	768	[2, 2, 18, 2]	[6, 12, 24, 48]	7

TABLE III
CONFIGURATIONS FOR ALL COMPARATIVE EXPERIMENTS

Methods	Source	Key parameters configurations
SRTHT [11]	PR'2021	Structure size: 12 × 12, Ring top-hat, M-estimator, local entropy
TLMS [12]	CEE'2021	Filter size: 15 × 15, Center Area:3 × 3
IPI [17]	TIP'2013	Patch Size: 50, sliding step: 10, $\lambda = 1/\sqrt{\max(m,n)}$, $\epsilon = 10^{-6}$
RIPT [16]	JSTARS' 2017	Patch Size: 50, sliding step:10, $\lambda = L/\sqrt{\min(I,J,P)}$, L=1, $\epsilon = 10^{-2}$, $\omega = 10^{-7}$
MPCM [14]	PR'2016	Patch scale size: 3, 5, 7, 9
MDvsFAcGAN [27]	ICCV'2019	Image size:128 × 128, $\lambda_1 = 100, \lambda_2 = 10$
ALCNet [34]	TGRS'2021	Image size:256 × 256, contextual scale:Global, Module:TLA-FPN
ACMNet [23]	WACV'2021	Image size:256 × 256, backbone:FPN, fuse mode:asymbi
IAANet [36]	TGRS'2022	Image size:128 × 128, Module:RPN
UIUNet [35]	TIP'2022	Image size:320 × 320, backbone:ResUnet, Module: RSU, IC-A
CGRANet [37]	JSTARS'2023	Image size:128 × 128,backbone:Res2Net,Module:CGM, MAB, RAM

TABLE IV
EVALUATION INDEX OF COMPARE EXPERIMENT

NUAA-SIRST	$IoU(\%)\uparrow$	$nIoU(\%)\uparrow$	$P_d(\%)\uparrow$	$F_a(10^{-6})\downarrow$
SRTHT [11]	23.46	22.12	79.51	23.31
TLMS [12]	24.11	23.18	80.21	30.1
IPI [17]	25.67	23.24	85.55	11.47
RIPT [16]	52.82	49.52	86.72	10.31
MPCM [14]	50.30	48.26	79.35	29.11
MDvsFAcGAN [27]	72.11	70.23	92.17	10.89
ALCNet [34]	73.69	71.19	97.01	12.21
ACMNet [23]	75.18	73.54	95.91	9.325
IAANet [36]	77.89	77.21	97.52	12.56
UIUNet [35]	78.89	78.21	98.55	8.324
CGRANet [37]	81.52	80.18	98.41	9.564
GSTUnet	82.61	80.97	98.91	5.32
NUDT-SIRST	$IoU(\%)\uparrow$	$nIoU(\%)\uparrow$	$P_d(\%)\uparrow$	$F_a(10^{-6})\downarrow$
SRTHT [11]	29.57	28.22	75.51	25.31
TLMS [12]	25.12	22.18	78.21	34.18
IPI [17]	27.76	24.98	75.28	13.68
RIPT [16]	67.76	64.21	82.28	12.68
MPCM [14]	64.16	62.69	77.55	19.87
MDvsFAcGAN [27]	73.11	72.23	94.17	14.89
ALCNet [34]	76.35	75.64	95.97	10.18
ACMNet [23]	81.15	78.92	96.54	9.231
IAANet [36]	78.81	77.93	98.54	14.56
UIUNet [35]	82.89	81.21	98.05	8.324
CGRANet [37]	83.52	82.18	98.09	9.564
GSTUnet	84.51	82.19	99.11	4.92

Display the best result in red font and the second-best result in blue, the up arrow means the larger is better, the down arrow means the less is better.

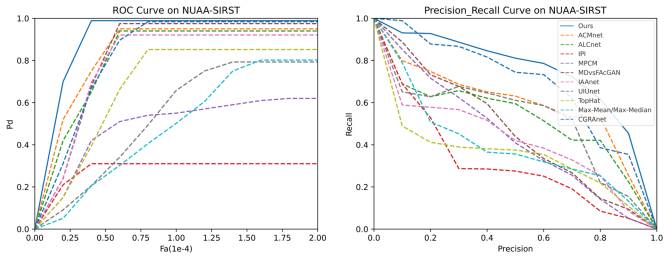

Fig. 10. ROC and Prec-Recall curves of different methods on the NUAA-SIRST dataset, Our method is blue solid line, other compared methods are dashed line.

larger AUC than filter-based, HVS-based, and LRM-based methods. Meanwhile, the method performs better than CNN-based methods, which proves the effectiveness of the proposed model.

b) *Qualitative results:* We visualized the results of the methods and 3-D gray figure on the NUAA-SIRST dataset with the comparison experiments, shown in Figs. 11–15. These experiments demonstrate that GSTUnet can accurately detect and locate targets even under low contrast and low SCRs conditions while obtains complete and precise target shapes. This high level of accuracy is achieved through the combination of GCL with Swin Transformer, which enables the effective establishment of a global view and the extraction of valuable edge information through the gated-shaped stream method and feature fusion modules.

Compared with traditional infrared small targets, which are susceptible to miss detection at low SCRs and FAs at high contrast. It is prone to have FAs and miss detection in the complex backgrounds, as shown in the yellow circle of MPCM, IPI in Figs. 11–15. Compared with traditional HVS, LRM-based methods, our method produces accurate target localization and shape outputs with very low FA rates. However, the traditional method performs well only on point-shaped targets, is less effective at characterizing shape, and is prone to localized highlighting that tends to produce many FAs, as shown the white circle in Figs. 11–15. While GSTUnet maintains high accuracy, the performance of traditional methods decreases dramatically when the size of the point target increases and the number

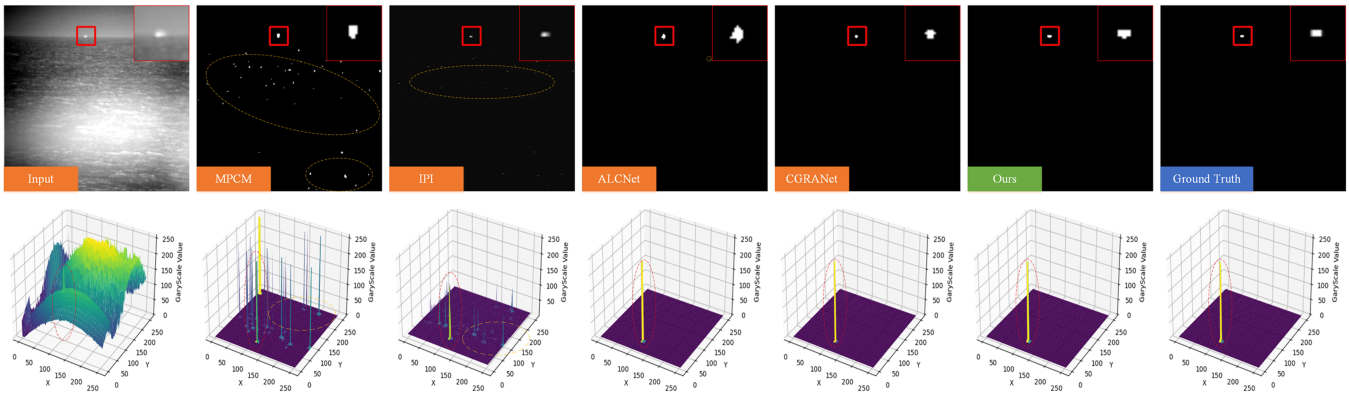

Fig. 11. Point small target on ocean backgrounds.

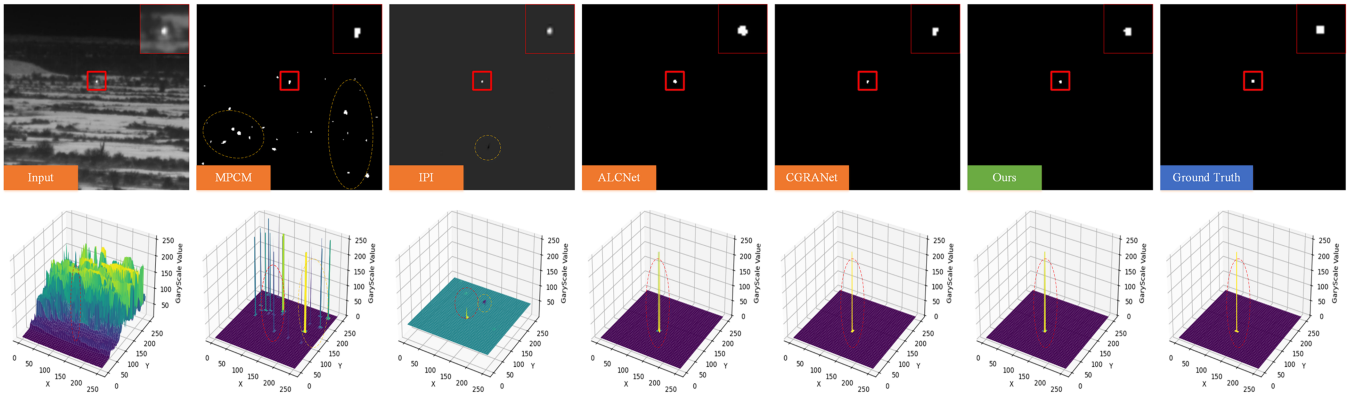

Fig. 12. Point small target on land backgrounds.

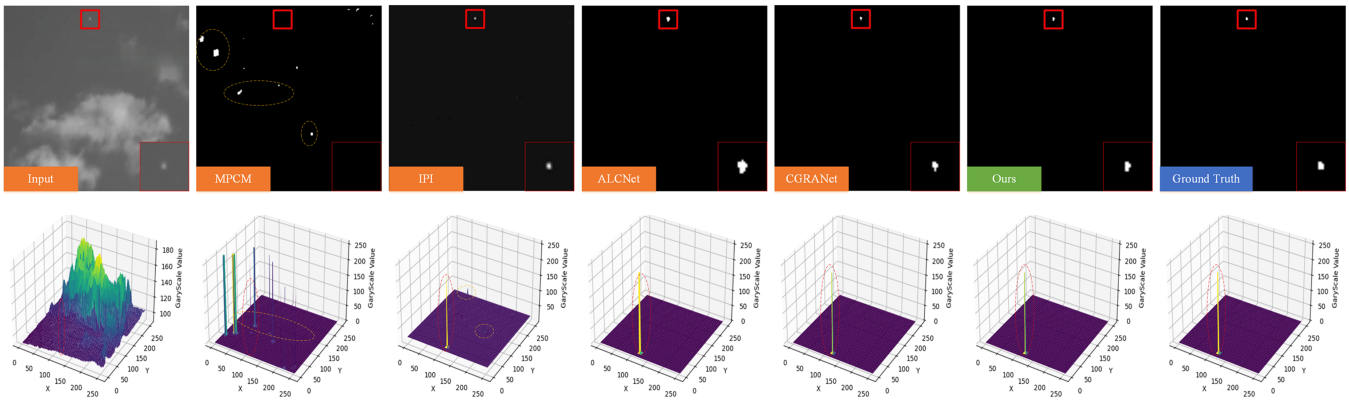

Fig. 13. Point target on sky backgrounds.

of bright spots with backgrounds interference increases. It is because the performance of the traditional methods is largely based on manual features and cannot be adapted to different backgrounds and various sizes.

CNN-based methods (i.e., CGRANet [37] and ALCNet [34]) perform much better than HVS-based and LRM-based methods, especially on the shape information of the small targets. However, due to the complexity of the scenes, many FAs and missed detection regions are generated, as shown by the white circles in Figs. 11–15. Our GSTUnet is more robust to these scene variations and can detect small targets on sky, ocean, and field backgrounds. In addition, our GSTUnet generates more accurate shape information than ALCNet [34], as shown in the zoomed region in Figs. 11–15. In summary, our GSTUnet can adapt to the challenges of different clutter backgrounds and target shapes, which achieves better performance.

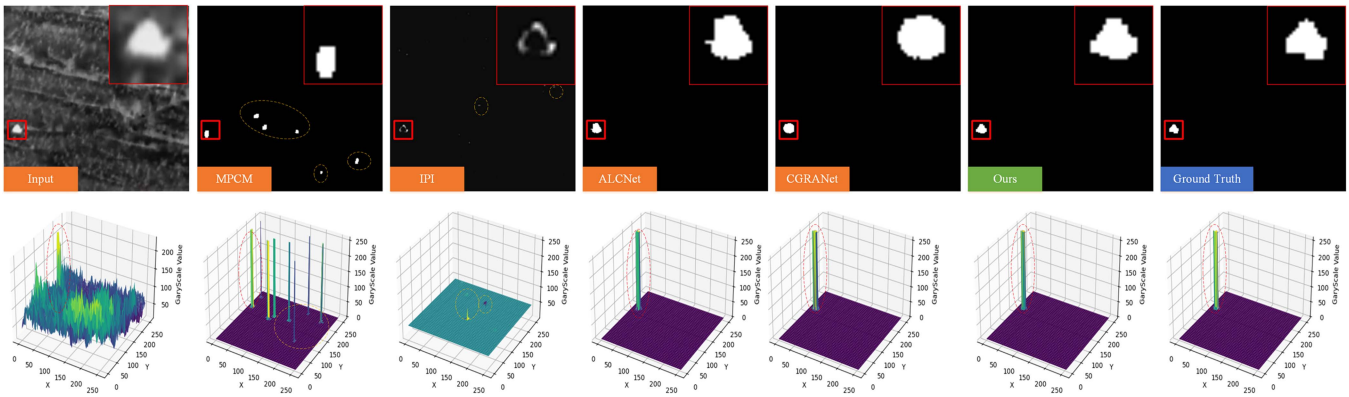

Fig. 14. Shape feature small target on land backgrounds.

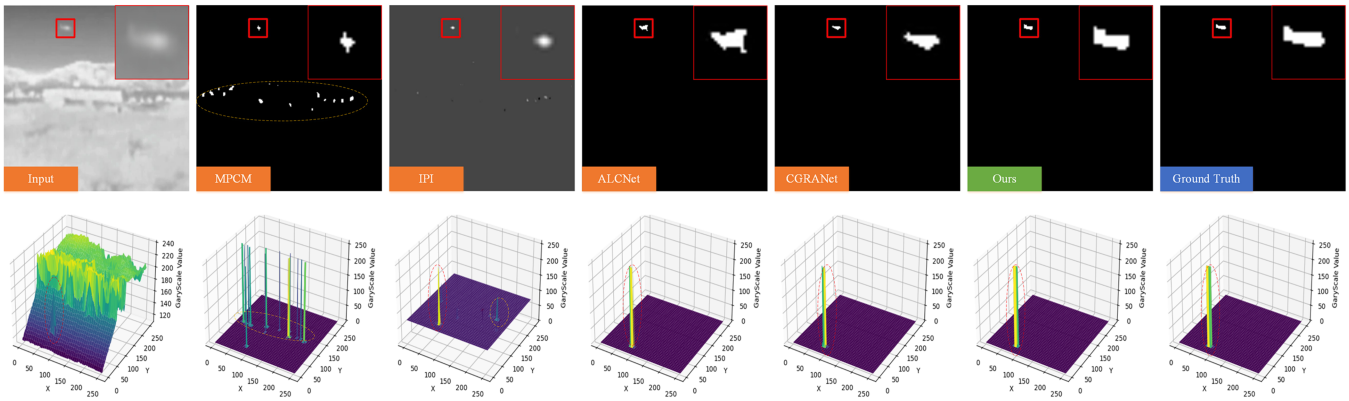

Fig. 15. Shape feature small target on sky backgrounds.

D. Ablation Study

To investigate the feasibility of each component of GSTUnet, we will take various ablation studies in this section, including utilizing the skip connection instead of the gated-shaped stream and exploring different scale structures Swin Transformer feature extractors' ability to the SISTD.

1) Impact of Gated-shaped Stream: To thoroughly investigate the effectiveness of ResNet in preserving small target features during gated-shaped stream, and the capability of the GCL module to effectively handle edge-related information in semantic features. The ablation study of the gated-shaped stream comprises two parts. The first part involves removing the second gated-shaped stream and replacing it with a skip connection between the encoder and decoder, as shown in Fig. 16. In addition, it is necessary to modify the loss function in Section III-D, which encompasses two components: DiceLoss and BCELoss, shown as follows:

$$\mathcal{L}_{\text{ablation}} = \mathcal{L}_{\text{Dice}}(y, X_{\text{out}}) + \lambda_a \mathcal{L}_{\text{BCE}}(y, X_{\text{out}}) \tag{14}$$

where y is the ground truth of the image, X_{out} is the output of the network, and λ_a is set to 1. The second study we remove the ResNet of gated-shaped stream, utilize only GCL. We plot the 3-D of the gray–scale map, enhancing the visualization results, clearly demonstrating the effectiveness of the residual

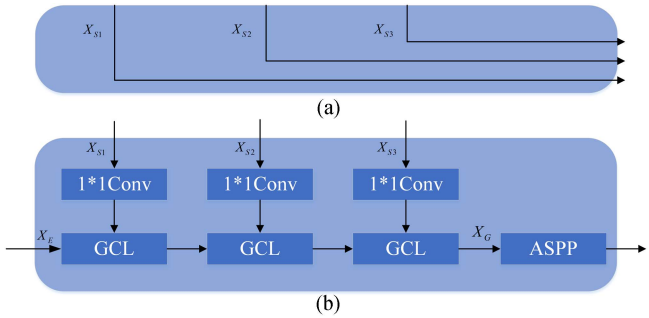

Fig. 16. Ablation study network structure on gated-shaped stream. (a) Skip connection between encoder and decoder. (b) Original gated-shaped stream remove ResNet18.

network in retaining small target features during downsampling as shown in Fig. 17. We can clearly observe that when there are targets of different scales, adding ResNet not only preserves the targets with shape features, but also preserves the features of small targets to overcome missed detections. After removing the gated-shaped stream, as shown in Fig. 17, facing the blurred edges of small targets, FAs are easily generated at the edges, and the shape information of small targets can not be well described.

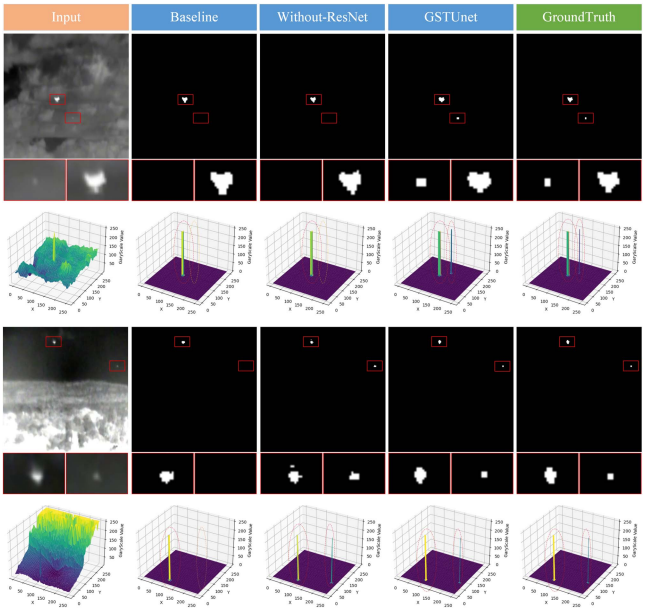

Fig. 17. Ablation study of baseline, GSTUnet and w/o ResNet18, the GSTUnet can accurately detect the edges and shapes of small targets, and also does not overlook small targets in complex backgrounds.

TABLE V
EVALUATION INDEX OF BASELINE AND GSTUNET

NUAA-SIRST		Metrics			
GCL	ResNet18	IoU(%) ↑	nIoU(%) ↑	P_d(%) ↑	$F_a(10^{-6})$ ↓
x	x	80.88	78.27	98.87	6.59
✓	x	81.38	79.71	98.92	5.89
✓	✓	**82.61**	**80.97**	**98.91**	**5.32**
NUDT-SIRST		Metrics			
GCL	ResNet18	IoU(%) ↑	nIoU(%) ↑	P_d(%) ↑	$F_a(10^{-6})$ ↓
x	x	82.07	80.79	98.68	5.82
✓	x	83.08	81.27	99.07	5.59
✓	✓	**84.51**	**82.19**	**99.11**	**4.92**

The bold values represents the best performing.

The quantitative results are shown in the Table V. After removing the gated-shaped stream, on the NUAA-SIRST dataset, the IoU, nIoU, and Pd decreased to 80.88%, 78.27%, 98.87%, and the FA increases to 6.59%. On the NUDT-SIRST dataset, the IoU, nIoU, and Pd decreased to 82.07%, 80.79%, 98.68%, and the FA increased to 5.82%. Subsequently, we plot the bar chart as Fig. 18, according to the quantitative result, the gated-shaped stream can approve the P_d and the IoU, and nIoU, decreasing F_a.

Meanwhile, we visualize the feature map of gated-shaped stream every 50 epochs, as shown in Fig. 19. As the number of training epoch increases, gated-shaped stream can learn local edge information of small targets.

2) Impact of Loss Function Weights: In order to explore the impact of different loss function weights on the overall performance of the network, we conducted ablation experiments on the weight of the loss function on (9). To investigate the impact of weights on the performance concerning small target edges and shapes, we conducted qualitative experiments. We

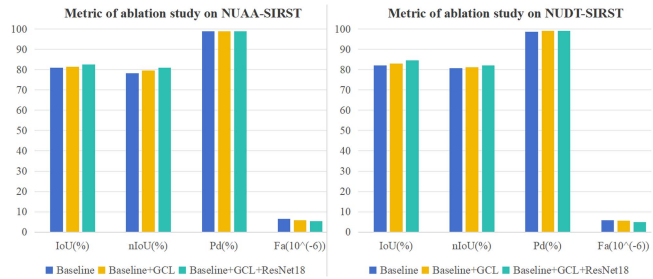

Fig. 18. Metric of ablation on two datasets, as the network increases, the various components we propose improve performance on Pd, IoU, and nIoU, whereas FA decreases.

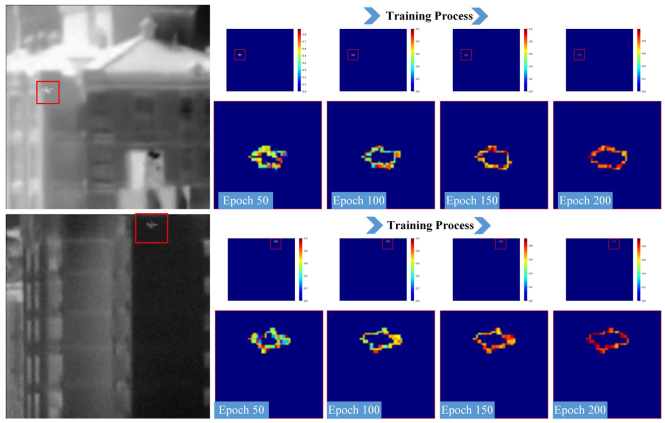

Fig. 19. Visualization of the output of gated-shaped stream during the training process with a heat map, we visualize every 50 epochs.

TABLE VI
EVALUATION INDEX OF DIFFERENT WEIGHTS OF THE LOSS FUNCTION

$\lambda_2 = 1, \lambda_3 = 0.5$		$\lambda_1 = 1, \lambda_3 = 0.5$		$\lambda_1 = 1, \lambda_2 = 1$	
λ_1	IoU	λ_2	IoU	λ_3	IoU
0.6	81.14	0.6	82.46	0.1	79.46
0.8	81.83	0.8	82.53	0.3	81.53
1.0	**82.61**	1.0	**82.61**	0.5	**82.61**
1.2	82.53	1.2	82.56	0.7	82.21
1.4	82.14	1.4	82.52	0.9	81.82

The bold values represents the best performing.

set $\lambda_1 = 1, \lambda_2 = 1,$ and $\lambda_3 = 0.5$ as the loss function weights in NUAA-SIRST to determine their effect on the IoU performance of the network (under consistent conditions as detailed in Table VI, training for 200 epochs and repeating the training ten times to calculate the average). The experimental results are presented in Table VI. Subsequently, we plotted the lines in three cases: $\lambda_1 = 1, \lambda_3 = 0.5$, $\lambda_1 = 1, \lambda_3 = 0.5$, and $\lambda_2 = 1, \lambda_3 = 0.5$, as shown in Fig. 20.

As shown in Fig. 20, changing the loss function influences the performance of the network, when λ_1 is set low, the network pays less attention to edge features, which will reduce the IoU. However, while the change of weights on loss function will only impact less on the overall performance, that also reflects the robustness of our method.

TABLE VII
EVALUATION INDEX OF SWIN TRANSFORMER BACKBONE VARIANTS

NUAA-SIRST	$IoU(\%)\uparrow$	$nIoU(\%)\uparrow$	$P_d(\%)\uparrow$	$F_a(10^{-6})\downarrow$
GSTUnet-T	82.48	80.61	98.87	5.59
GSTUnet-B	82.58	80.27	98.47	5.41
GSTUnet-L	**82.61**	**80.97**	**98.91**	**5.32**
NUDT-SIRST	$IoU(\%)\uparrow$	$nIoU(\%)\uparrow$	$P_d(\%)\uparrow$	$F_a(10^{-6})\downarrow$
GSTUnet-T	84.07	82.29	99.06	5.22
GSTUnet-B	84.52	82.02	99.96	5.01
GSTUnet-L	**84.51**	**82.19**	**99.11**	**4.92**

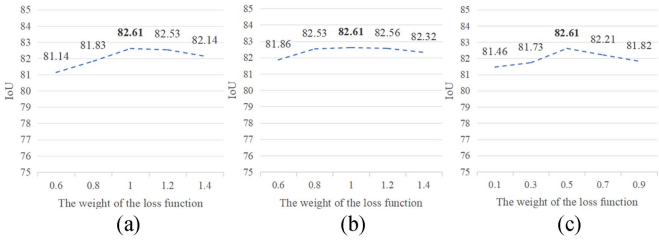

Fig. 20. Change tendency of different weights of loss function.

3) Impact of Swin Transformer Backbone: To explore the different parameters of the feature extractor, we replaced the pretrained model weights with the Swin-T, Swin-B, and Swin-L feature extraction backbone. The various parameters are shown in Table II. The parameter configuration settings are the same as those in Table I.

According to the ablation study. The result of different backbone are shown in Table VII three different sizes of Swin Transformer backbones have little impact on infrared small object detection tasks. Swin-L has the largest number of parameters but it is not significantly ahead of Swin-T and Swin-B in terms of evaluation index. It is because infrared images have low resolution, which makes large backbones unable to leverage their global vision capabilities fully. It also indicates that lightweight backbones can achieve good results and have broad application prospects, which provides potential value for model light-weighting.

V. CONCLUSION

This work proposes the GSTUnet, a network with two data streams based on U-shaped encoder-decoder architecture for infrared small object detection. Our GSTUnet is built upon a hierarchical Swin Transformer and innovatively incorporates Swin Transformer modules in the decoder. In addition, we introduce a gated-shaped stream based on the GCL after feature extraction, which focuses solely on extracting and computing edge information of small infrared objects. We have trained the neural network end-to-end and performed an upsampling information fusion from the two data streams by adding the edge loss and overall loss function. We establish long-term dependency relationships between different scale features through a self-attention mechanism and effectively fuse multiscale features with edge-sensitive results. Extensive experiments on real and synthetic datasets demonstrate that our GSTUnet significantly outperforms other advanced methods. In future work, we will focus on designing lighter transformer-based models and achieving better pixel-level intrinsic structural features generated by patch partitioning in visual transformers.

REFERENCES

[1] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Small infrared target detection based on weighted local difference measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4204–4214, Jul. 2016.
[2] R. Kou et al., "Infrared small target segmentation networks: A survey," *Pattern Recognit.*, vol. 143, 2023, Art. no. 109788.
[3] Y. Han, J. Liao, T. Lu, T. Pu, and Z. Peng, "KCPNet: Knowledge-driven context perception networks for ship detection in infrared imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5000219.
[4] Y. Han, X. Yang, T. Pu, and Z. Peng, "Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5612318.
[5] Y. Sun, J. Yang, and W. An, "Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3737–3752, May 2021.
[6] Z. Qiu, Y. Ma, F. Fan, J. Huang, and L. Wu, "Global sparsity-weighted local contrast measure for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7507405.
[7] Z. Qiu, Y. Ma, F. Fan, J. Huang, and M. Wu, "Adaptive scale patch-based contrast measure for dim and small infrared target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 7000305.
[8] Y. Dai, X. Li, F. Zhou, Y. Qian, Y. Chen, and J. Yang, "One-stage cascade refinement networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5000917.
[9] X. Bai and F. Zhou, "Analysis of new top-hat transformation and the application for infrared dim small target detection," *Pattern Recognit.*, vol. 43, no. 6, pp. 2145–2156, 2010.
[10] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," *SPIE*, vol. 3809, pp. 74–83, 1999.
[11] L. Deng, J. Zhang, G. Xu, and H. Zhu, "Infrared small target detection via adaptive m-estimator ring top-hat transformation," *Pattern Recognit.*, vol. 112, 2021, Art. no. 107729.
[12] H. Li, Q. Wang, H. Wang, and W. K. Yang, "Infrared small target detection using tensor based least mean square," *Comput. Elect. Eng.*, vol. 91, 2021, Art. no. 106994.
[13] C. L. Philip Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A. local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.
[14] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, 2016.
[15] Z.-B. Qiu, Y. Ma, F. Fan, J. Huang, M.-H. Wu, and X.-G. Mei, "A pixel-level local contrast measure for infrared small target detection," *Defence Technol.*, vol. 18, no. 9, pp. 1589–1601, 2022.
[16] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both non-local and local priors for single-frame small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017.
[17] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.
[18] L. Zhang, L. Peng, T. Zhang, S. Cao, and Z. Peng, "Infrared small target detection via non-convex rank approximation minimization joint l 2, 1 norm," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1821.
[19] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 382.
[20] F. Yan, G. Xu, Q. Wu, J. Wang, and Z. Li, "Infrared small target detection using kernel low-rank approximation and regularization terms for constraints," *Infrared Phys. Technol.*, vol. 125, 2022, Art. no. 104222.
[21] M. Liu, H.-Y. Du, Y.-J. Zhao, L.-Q. Dong, M. Hui, and S. X. Wang, "Image small target detection based on deep learning with SNR controlled sample generation," *Curr. Trends Comput. Sci. Mech. Automat.*, vol. 1, pp. 211–220, 2017.

[22] B. McIntosh, S. Venkataramanan, and A. Mahalanobis, "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.*, vol. 57, no. 1, pp. 485–496, Feb. 2021.
[23] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 950–959.
[24] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 877–886.
[25] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2022.
[26] J. Lin, K. Zhang, X. Yang, X. Cheng, and C. Li, "Infrared dim and small target detection based on U-Transformer," *J. Vis. Commun. Image Representation*, vol. 89, 2022, Art. no. 103684.
[27] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8509–8518.
[28] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 4005615.
[29] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[31] F. Wu, H. Yu, A. Liu, J. Luo, and Z. Peng, "Infrared small target detection using spatio-temporal 4 d tensor train and ring unfolding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5002922.
[32] M. Ju, J. Luo, G. Liu, and H. Luo, "ISTDet: An efficient end-to-end neural network for infrared small target detection," *Infrared Phys. Technol.*, vol. 114, 2021, Art. no. 103659.
[33] K. Qian, S.-J. Zhang, H.-Y. Ma, and W.-J. Sun, "SiamIST: Infrared small target tracking based on an improved SiamRPN," *Infrared Phys. Technol.*, vol. 134, 2023, Art. no. 104920.
[34] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9813–9824, Nov. 2021.
[35] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.
[36] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002013.
[37] S. Zhong, F. Zhang, and J. Duan, "Context guided reverse attention network with multiscale aggregation for infrared small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9725–9734, 2023.
[38] T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-guided pyramid context networks for detecting infrared small target under complex background," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 4, pp. 4250–4261, Aug. 2023.
[39] R. Kou, C. Wang, Y. Yu, Z. Peng, F. Huang, and Q. Fu, "Infrared small target tracking algorithm via segmentation network and multi-strategy fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612912.
[40] R. Kou et al., "LW-IRSTNet: Lightweight infrared small target segmentation network and application deployment," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5621313.
[41] F. Lin, S. Ge, K. Bao, C. Yan, and D. Zeng, "Learning shape-biased representations for infrared small target detection," *IEEE Trans. Multimedia*, vol. 26, pp. 4681–4692, 2023.
[42] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 6000–6010, 2017.
[43] L. Zhang and Y. Wen, "A transformer-based framework for automatic Covid19 diagnosis in chest CTS," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 513–518.
[44] S. Hao, B. Wu, K. Zhao, Y. Ye, and W. Wang, "Two-stream Swin Transformer with differentiable Sobel operator for remote sensing image classification," *Remote Sens.*, vol. 14, no. 6, 2022, Art. no. 1507.
[45] L. Gao, J. Zhang, C. Yang, and Y. Zhou, "Cas-VSwin transformer: A variant Swin Transformer for surface-defect detection," *Comput. Ind.*, vol. 140, 2022, Art. no. 103689.
[46] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Comput. Vis.–ECCV 16th Eur. Conf.*, 2020, pp. 213–229.
[47] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[48] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Comput. Vis.–ECCV Workshops*, 2022, pp. 205–218.
[49] C. Li, Z. Huang, X. Xie, and W. Li, "IST-Transnet: Infrared small target detection based on transformer network," *Infrared Phys. Technol.*, vol. 132, 2023, Art. no. 104723.
[50] F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, "Infrared small and dim target detection with transformer under complex backgrounds," *IEEE Trans. Image Process.*, vol. 32, pp. 5921–5932, 2023.
[51] Y. Zhu, S. Tang, Y. Jiang, and R. Kang, "Dau-Net: A regression cell counting method," in *Proc. 6th Int. Conf. Inf. Sci., Comput. Technol. Transp.*, 2021, pp. 1–6.
[52] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[53] K. R. Castleman. *Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice Hall Press, 1996.
[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[55] L. Yuan and X. Xu, "Adaptive image edge detection algorithm based on canny operator," in *Proc. 4th Int. Conf. Adv. Inf. Technol. Sensor Appl.*, 2015, pp. 28–31.
[56] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support: 3rd Int. Workshop, DLMIA 2017, 7th Int. Workshop, ML-CDS 2017, Held Conjunction*, 2017, pp. 240–248.

Yiming Zhu received the master's degree in optical engineering from the School of Optics and Photonics, Beijing Institute of Technology (BIT), Beijing, China, in 2022. He is currently working toward the doctoral degree with the Multispectral Vision Processing (MVP) Laboratory, the School of Electronic Information, Wuhan University, Wuhan, China.

His research interests include computer vision, infrared technology, and remote sensing.

Yong Ma received the B.S. degree in industrial automation and the M.S. degree in automatic control from the Beijing Institute of Technology, Beijing, China, in 1994 and 1997, respectively, and the Ph.D. degree in electronic circuit and system from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003.

Between 2004 and 2006, he was a Lecturer with the University of the West of England, Bristol, U.K. From 2006 and 2014, he was with the Wuhan National Laboratory for Optoelectronics, HUST, Wuhan, where he was a Professor of electronics. He is currently a Professor with the Electronic Information School, Wuhan University, Wuhan. His research interests include signal and systems, infrared image processing, pattern recognition, and interface circuits to sensors and actuators.

Fan Fan (Member, IEEE) received the B.S. degree in communication engineering and the Ph.D. degree in electronic circuit and system from the Huazhong University of Science and Technology, Wuhan, China, in 2009, and 2015, respectively.

He is currently an Associate Professor with the Electronic Information School, Wuhan University, Wuhan. His research interests include infrared thermal imaging, machine learning, and computer vision.

Jun Huang received the B.S. degree in communication engineering and the Ph.D. degree in electronic circuit and system from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively.

He is currently an Associate Professor with the Electronic Information School, Wuhan University, Wuhan. His main research interests include infrared image and infrared spectrum processing.

Kangle Wu received the B.S. and M.S. degrees in automation and control engineering from the School of Automation, China University of Geosciences, Wuhan, China, in 2019 and 2022, respectively. He is currently working toward the Ph.D. degree with the Electronic Information School, Wuhan University, Wuhan.

His research interests include neural networks, machine learning, and image fusion.

Ge Wang received the B.S. degree in communication engineering and the M.S. degree in electronics and communication engineering from Hubei University, Wuhan, China, in 2019 and 2022, respectively. He is currently working toward the Ph.D. degree in electronic information with the School of Electronic Information, Wuhan University, Wuhan.

His research interests include image segmentation, image fusion, and smart agriculture.