SCTransNet: Spatial-Channel Cross Transformer Network for Infrared Small Target Detection

Shuai Yuan, *Student Member, IEEE*, Hanlin Qin, *Member, IEEE*, Xiang Yan, *Member, IEEE*, Naveed Akhtar, *Member, IEEE*, and Ajmal Mian, *Senior Member, IEEE*

***Abstract*— Infrared small target detection (IRSTD) has recently benefitted greatly from U-shaped neural models. However, largely overlooking effective global information modeling, existing techniques struggle when the target has high similarities with the background. We present a *S*patial-channel *C*ross *T*ransformer *Net*work (SCTransNet) that leverages spatial-channel cross transformer blocks (SCTBs) on top of long-range skip connections (SKs) to address the aforementioned challenge. In the proposed SCTBs, the outputs of all encoders are interacted with cross transformer to generate mixed features, which are redistributed to all decoders to effectively reinforce semantic differences between the target and clutter at full levels. Specifically, SCTB contains the following two key elements: 1) spatial-embedded single-head channel cross-attention (SSCA) for exchanging local spatial features and full-level global channel information to eliminate ambiguity among the encoders and facilitate high-level semantic associations of the images and 2) a complementary feed-forward network (CFN) for enhancing the feature discriminability via a multiscale strategy and cross-spatial-channel information interaction to promote beneficial information transfer. Our SCTransNet effectively encodes the semantic differences between targets and backgrounds to boost its internal representation for detecting small infrared targets accurately. Extensive experiments on three public datasets, NUDT-SIRST, NUAA-SIRST, and IRSTD-1K, demonstrate that the proposed SCTransNet outperforms existing IRSTD methods. Our code will be made public at https://github.com/xdFai/SCTransNet.**

Manuscript received 29 February 2024; revised 22 March 2024; accepted 27 March 2024. Date of publication 1 April 2024; date of current version 12 April 2024. This work was supported in part by Shaanxi Province Key Research and Development Plan Project under Grant 2022JBGS2-09; in part by the 111 Project under Grant B17035; in part by Shaanxi Province Science and Technology Plan Project under Grant 2023KXJ-170; in part by Xi'an City Science and Technology Plan Project under Grant 21JBGSZ-QCY9-0004, Grant 23ZDCYJSGG0011-2023, Grant 22JBGS-QCY4-0006, and Grant 23GBGS0001; in part by the Aeronautical Science Foundation of China under Grant 20230024081027; in part by the Natural Science Foundation Explore of Zhejiang Province under Grant LTGG24F010001; in part by the Natural Science Foundation of Ningbo under Grant 2022J185; in part by China Scholarship Council under Grant 202306960052; in part by the Technology Area Foundation of China under Grant 2021-JJ-1244, Grant 2021-JJ-0471, and Grant 2023-JJ-0148; and in part by Xidian Graduate Student Innovation fund under Grant YJSJ23010. *(Corresponding authors: Hanlin Qin; Xiang Yan.)*

Shuai Yuan, Hanlin Qin, and Xiang Yan are with the School of Optoelectronic Engineering, Xidian University, Xi'an 710071, China (e-mail: yuansy@stu.xidian.edu.cn; hlqin@mail.xidian.edu.cn; xyan@xidian.edu.cn).

Naveed Akhtar is with the School of Computing and Information Systems, Faculty of Engineering and IT, The University of Melbourne, Parkville, VIC 3052, Australia (e-mail: naveed.akhtar1@unimelb.edu.au).

Ajmal Mian is with the Department of Computer Science and Software Engineering, The University of Western Australia, Perth, WA 6009, Australia (e-mail: ajmal.mian@uwa.edu.au).

Digital Object Identifier 10.1109/TGRS.2024.3383649

***Index Terms*— Convolutional neural network (CNN), cross-attention, deep learning, infrared small target detection (IRSTD), transformer.**

I. Introduction

INFRARED small target detection (IRSTD) plays an important role in traffic monitoring [1], maritime rescue [2], and target warning [3], where separating small targets in complex scene backgrounds is required. The challenges emerging from the dynamic nature of scenes have attracted considerable research attention in single-frame IRSTD [4]. Early methods in this direction employed image filtering [5], [6], human visual system (HVS) [7], [8], and low-rank approximation [9], [10] techniques while relying on complex handcrafted feature designs, empirical observations, and model parameter fine-tuning. However, suffering from the absence of a reliable high-level understanding of the holistic scene, these methods exhibit poor robustness.

Recently, learning-based methods have become more popular due to their strong data-driven feature mining abilities [11]. To capture the target's outlines and mitigate performance degradation caused by its small size, these methods approach the IRSTD problem as a semantic segmentation task instead of a traditional object detection issue. Unlike general object segmentation in autonomous driving [12], imaging mechanism of the IR detection systems in remote sensing applications [13] leads to small targets in images exhibiting the following characteristics.

1) Dim and small: Due to remote imaging, IR targets are small and usually exhibit a low signal-to-clutter ratio, making them susceptible to immersion in heavy noise and background clutter.
2) Characterless: Thermal images lack color and texture information in targets, and imprecise camera focus can cause target blurring. These factors pose peculiar challenges in designing feature extraction techniques for IRSTD.
3) Uncertain shapes: The scales and shapes of IR targets vary significantly across different scenes, which makes the problem of detection considerably challenging.

To identify small IR targets in complex backgrounds, numerous learning-based methods have been proposed, among which neural networks with U-shaped architectures have gained prominence. Benefiting from these frameworks of

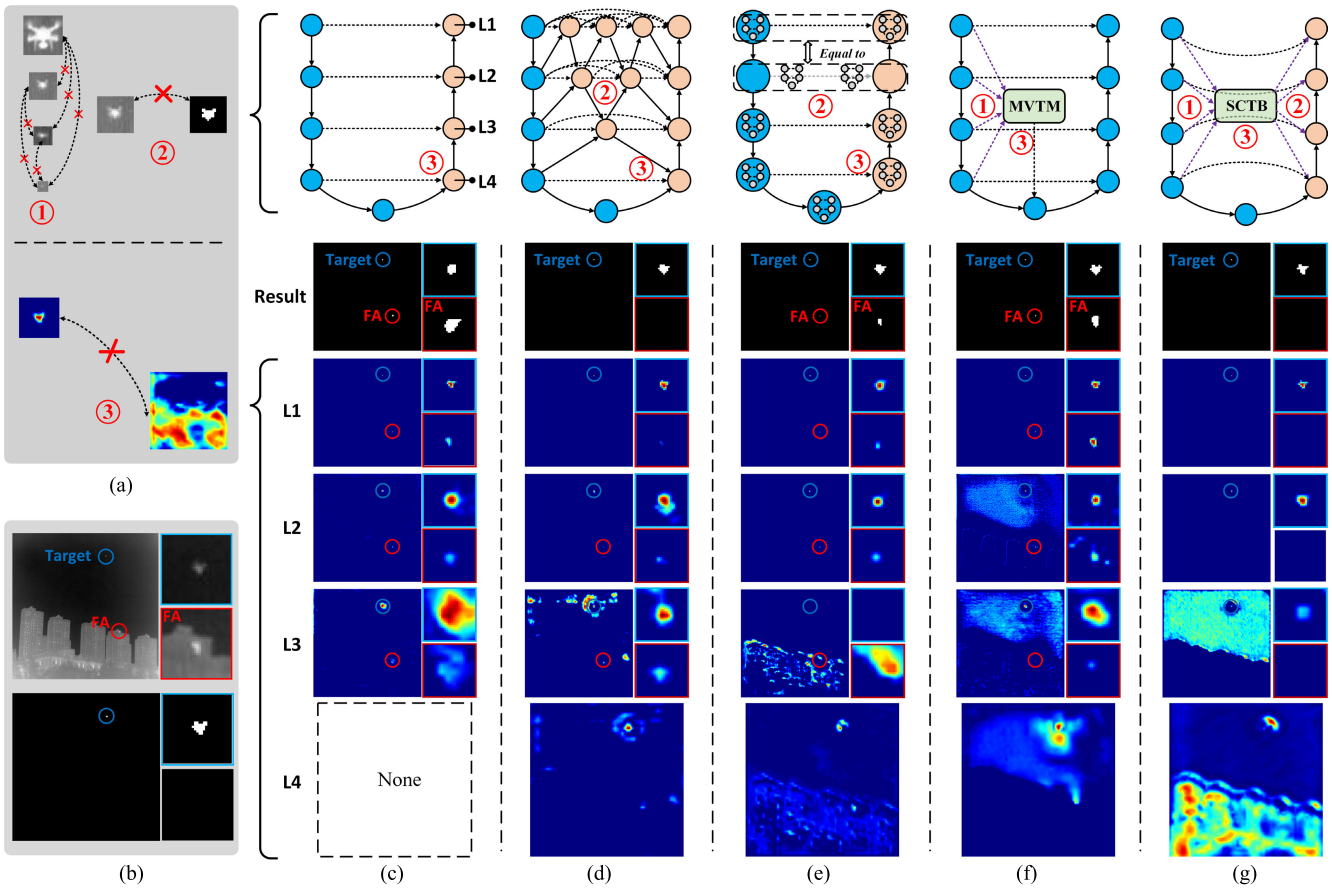


Fig. 1. Framework and visualization maps of representative IRSTD methods, with each method's frame labeled according to the specific challenge it addresses. The visualization maps show that the CNN-based approaches (ACM [14], DNA-Net [15], and UIU-Net [16]) focus on modeling the local information of the target and less on establishing the global semantic information of the image; mixer CNN and transformer methods (MTU-Net [17] and our SCTransNet) pay more attention to the background global information and target semantics. Only our method meticulously models buildings and the sky separately in the high-level semantic map, effectively distinguishing the target from the background, reducing false alarms. (a) Three challenges. (b) Input image and GT. (c) ACM. (d) DNA-Net. (e) UIU-Net. (f) MTU-Net. (g) SCTransNet.

encoders, decoders, and long-range SKs, asymmetric contextual modulation (ACM) network [14] initially demonstrated the effectiveness of cross-layer feature fusion for retaining IR target features. This is achieved through bidirectional aggregation of high-level semantic information and low-level details using asymmetric top-down and bottom-up structures. Subsequently, feature fusion strategies have been widely adopted in IRSTD task [18], [19], [20], [21]. A few recent methods facilitate the transfer of beneficial features to the decoder component by improving the SKs [22], [23]. Inspired by the nested structure [24], DNA-Net [15] developed a densely nested interactive module to facilitate gradual interaction between high- and low-level features and adaptively enhance features. Moreover, there are also approaches that focus on developing more effective encoders and decoders [25], [26]. For instance, UIU-Net [16] embeds smaller U-Nets in the U-Net to learn the local contrast information of the target and perform interactive cross-attention (IC-A) for feature fusion.

Despite achieving satisfactory results, the aforementioned convolutional neural network (CNN)-based approaches lack the ability to encode comprehensive attributes of the target, missing their discriminative features. To address that, MTU-Net [17] employs a multilevel vision transformer (ViT)-CNN hybrid encoder to exploit the spatial correlation among all encoded features for contextual information aggregation. However, a simple spatial ViT-CNN hybrid module is insufficient for understanding the global semantics of images, which makes high false alarms. To further dissect the issue, we illustrate the frameworks of ACM [14], DNA-Net [15], UIU-Net [16], and MTU-Net [17] separately, along with visualizations of the attention maps from different decoder levels in Fig. 1(c)–(f). Given the input image in Fig. 1(b), we observe that false alarms occur when existing models direct their attention to localized regions of background clutters in high-level features. In other words, false alarms are often caused by discontinuity modeling of backgrounds in the deeper layers. We identify this problem to the following three main reasons.

A. Semantic Interaction Across Feature Levels Is Not Established Well

As shown in Fig. 1(a)-①, IR small targets exhibit limited features owing to their diminutive size. Multiple downsampling processes inevitably result in the loss of spatial information. This considerably affects the level-to-level feature interactions in the network, eventually leading to poor comprehensive global semantic information encoding.

B. Feature Enhancement Fails to Bridge the Information Gap Between Encoders and Decoders

As shown in Fig. 1(a)-②, there exists a semantic gap between the output features of encoders and the input features of the decoders. Simple SKs and dense nested modules are insufficient to enhance the advantageous responses of the features to the decoder, thereby making it challenging to establish a mapping relationship from the IR image to the segmentation space.

C. Inaccurate Long-Range Contextual Perception of Targets and Backgrounds in Deeper Layers

IR small targets can be highly similar to the scene background. As shown in Fig. 1(a)-③, a powerful detector not only has to sense the local saliency of the target but also needs to model the continuity of the background. CNNs and vanilla ViTs are not fully equipped to achieve this.

Inspired by the success of channel-wise cross-fusion transformer (CCT) in image segmentation [27], [28], [29] and local spatial embedding in image restoration [30], [31], [32], we propose a spatial-channel cross transformer network (SCTransNet) for IRSTD to address the above challenges, aiming to distinguish the small targets and background clutters in deeper layers. As illustrated in Fig. 1(g), our framework adds multiple spatial-channel cross transformer blocks (SCTBs) (Section III-B) on the original SKs to establish an explicit association with all encoders and decoders. Specifically, SCTB consists of two components: spatial-embedded single-head channel cross-attention (SSCA) (Section III-B1) and complementary feed-forward network (CFN) (Section III-B2).

The SSCA applies channel cross-attention from the feature dimension at all levels to learn global information. Besides, depth-wise convolutions are used for local spatial context mixing before feature covariance computation. This strategy provides two advantages. First, it highlights the context of local space with a small computational overhead using the convolution's local connectivity, thereby increasing the saliency of IR small targets. Second, it makes sure that contextualized global relationships among full-level feature pixels are implicitly captured during the attention matrix computation, thereby reinforcing the continuity of the background.

After the SSCA completes the cross-level information interaction, CFN performs feature enhancement at every level in two complementary stages. Initially, it utilizes multiscale depth-wise convolutions to enhance target neighborhood space response and pixel-wise aggregates the cross-channel nonlinear information. Subsequently, it estimates total spatial information on a channel-by-channel basis using global average pooling (GAP) and creates local cross-channel interactions between distinct semantic patterns as an attention map. The above strategy has two advantages. 1) Multiscale spatial modeling can emphasize semantic differences between the target and background. 2) Establishing the complementary correlation of the local space global channel (LSGC) and the global space local channel (GSLC) can facilitate the interface between infrared images and semantic maps.

Benefiting from the above structure [Fig. 1(g)], our SCTransNet can perceive the image semantics better than other methods leading to reduced false alarms. Our main contributions are as follows.

1) We propose SCTransNet, which leverages multiple SCTBs connecting all encoders and decoders to predict the context of targets and backgrounds in the deeper network layers.
2) We propose an SSCA module to foster semantic interactions across all feature levels and learn the long-range context correlation of the image.
3) We devise a novel CFN by crossing spatial-channel information to enhance the semantic difference between the target and background, bridging the semantic gap between encoders and decoders.

II. Related Work

We first briefly review the CNN- and transformer-based techniques in IRSTD. Following that, we discuss the application of channel-wise cross transformer in image processing.

A. CNN-Based IRSTD Methods

Owing to the local saliency of IR small targets coinciding with the local connectivity of CNNs, CNNs have demonstrated remarkable performance in the IRSTD task. To effectively preserve the semantic patterns of small targets, diverse feature fusion strategies have been proposed. One common strategy is cross-layer feature fusion [33], [34], [35], which can address the loss of target information when fusing the encoded and decoded features. Additionally, densely nested interactive feature fusion [15], [36] is used to repetitively fuse and enhance the features of different levels, maintaining the information of IR small targets in the deeper layers. Considering variations in target scales, multiscale feature fusion [37], [38] has been proposed to enhance the low-resolution feature maps. Besides feature fusion, incorporating prior information about the target into CNNs is also an effective strategy. For instance, Sun et al. [39] exploited the small-target gray gradient change property using a receptive-field and direction-induced attention network (RDIAN), which solves the imbalance between the target and background classes. Zhang et al. [40] used Taylor's finite difference for complex edge feature extraction of a target to enhance the target and background gray scale difference.

Although satisfactory results are achieved by CNN-based techniques, the inherent inductive bias of CNNs makes it difficult to unambiguously establish long-range contextual information for the IRSTD task. Unlike the aforementioned methods, we incorporate transformer blocks into the backbone of CNNs as a core unit to capture non-local information for the entire image.

B. Transformer-Based IRSTD Methods

ViT [41] decomposes an image/features into a series of patches and computes their correlation. This computational paradigm can stably establish long-distance dependence among different patches, leading to its widespread usage in IRSTD tasks for global image modeling [42], [43], [44]. Inspired by TransUnet [45], IRSTFormer [46] embedded the

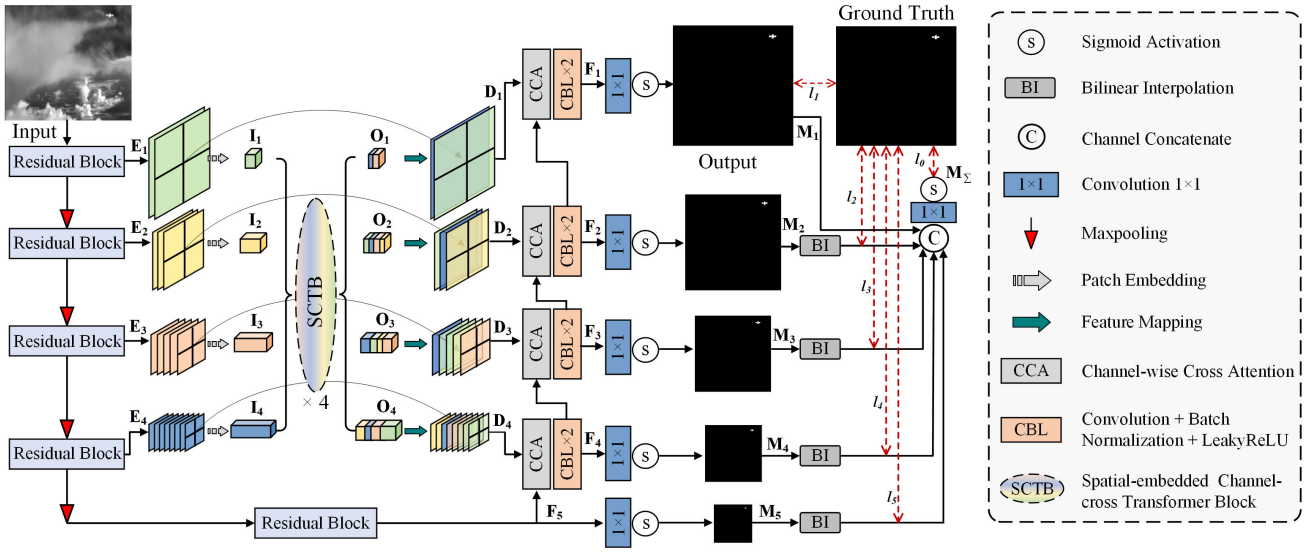


Fig. 2. Overview of the proposed SCTransNet for infrared small object detection. Our SCTransNet adopts a U-shaped structure and adds four SCTBs on the long-range SKs, and the multiscale deeply supervised fusion strategy is used to optimize our SCTransNet.

spatial transformer within multiple encoder stages in a U-Net. Motivated by Swin transformer [47], FTC-Net [48] establishes a robust feature representation of the target using a two-branch structure combining the local feature extraction of CNNs and the global feature extraction capability of the Swin transformer. Recently, Meng et al. [49] modeled the local gradient information of the target using central difference convolution and employed criss-cross multiattention [12] to acquire contextual information. Note that, the above methods use spatial self-attention (SA) to calculate covariance-based attention maps, which have two problems: 1) the computational complexity is proportional to the square of the number of tokens, which limits the multiple nesting of the spatial transformer and its fine-grained representation of high-resolution images [30] and 2) the SA only constructs long-distance dependency for a single feature map, whereas it is more critical to establish contextual connections among all levels.

Different from previous works, we present the channel-wise cross transformer on the long-range SKs for the first time in the IRSTD task. This allows establishing cross-channel semantic patterns across all levels with an acceptable computational overhead.

C. Channel-Wise Cross Transformer on Image Processing

Unlike spatial transformers, channel-wise transformers (CTs) [30] treat each channel as a patch. Note that every channel is a unique semantic pattern, CT essentially establishes correlations between multiple semantic patterns. Considering that not every SK is effective, Wang et al. [27] proposed UCTransNet, utilizing CCT to address the semantic difference for precise medical image segmentation. The CCT's powerful global semantic modeling capability facilitates its widespread application in tasks such as metal surface defect detection [29], remote sensing image segmentation [50], and building edge detection [28]. This inspires us to introduce this model to separate IR targets and backgrounds in the deeper layers effectively. However, IR small targets differ significantly from the usual large-size targets not only in size but also in terms of effective features and sample balance. The attention matrix computation, the positional encoding, and the pure channel modeling in vanilla CCT are harmful to the limited-pixel target detection. Therefore, we propose an SCTB. Its launching point is leveraging the target's local spatial saliency and global background continuity to separate the target in the deep layers.

III. Method

This section elaborates on the proposed SCTransNet for IRSTD. We begin by presenting the overall structure of the proposed SCTransNet in Section III-A. Then, we present the technical details of the SCTB and its internal structure: SSCA and the CFN in Section III-B.

A. Overall Pipeline

As shown in Fig. 2, given an infrared image, SCTransNet initially employs four groups of residual blocks (RBs) [51] and max-pooling layers, to acquire high-level features $\mathbf{E_i} \in \mathbb{R}^{C_i \times (H/i) \times (W/i)}$, $(i = 1, 2, 3, 4)$. C_i are the channel dimensions, in which $C_1 = 32$, $C_2 = 64$, $C_3 = 128$, and $C_4 = 256$. Next, we perform patch embedding on $\mathbf{E_i}$ using convolution with kernel size and stride size of P, $P/2$, $P/4$, and $P/8$ to obtain embedded layers $\mathbf{I_i} \in \mathbb{R}^{C_i \times (H/16) \times (W/16)}$, $(i = 1, 2, 3, 4)$, respectively. These layers are then fed into the SCTB for full-level semantic feature blending and obtaining the output $\mathbf{O_i} \in \mathbb{R}^{C_i \times (H/16) \times (W/16)}$, $(i = 1, 2, 3, 4)$, which have the same size of $\mathbf{I_i}$. Details of SCTB are provided in Section III-B. $\mathbf{O_i}$ are recovered to the size of the original encoder processing using feature mapping (FM), which consists of bilinear interpolation, convolution, batch normalization, and ReLU activation. Meanwhile, we employ a residual connection to merge the features between the encoders and decoders. The process described above can be expressed mathematically as

$$\mathbf{O_i} = \mathbf{E_i} + \mathrm{FM}_i(\mathrm{SCTB}(\mathbf{I_1}, \mathbf{I_2}, \mathbf{I_3}, \mathbf{I_4})) \ (i = 1, 2, 3, 4). \quad (1)$$

Finally, the channel-wise cross-attention (CCA) [27] is employed to fuse the high- and low-level features, followed by decoding using two CBL blocks.

To enhance the gradient propagation efficiency and feature representation, we utilize a multiscale deeply supervised fusion strategy to optimize SCTransNet. Specifically, a 1×1 convolution and sigmoid function are used for each decoder outputs $\mathbf{F_i}$, acquiring the saliency map $\mathbf{M_i}$ which is denoted as

$$\mathbf{M_i} = \text{Sigmoid}(f_{1\times1}(\mathbf{F_i}))\ (i = 1, 2, 3, 4, 5). \tag{2}$$

Next, we upsample the low-resolution salient maps $\mathbf{M_i}$ $(i = 2, 3, 4, 5)$ to the original image size and fuse all the salient maps to obtain $\mathbf{M}_{\sum}$ as

$$\mathbf{M}_{\sum} = \text{Sigmoid}(f_{1\times1}[\mathbf{M_1}, \mathcal{B}(\mathbf{M_2}), \mathcal{B}(\mathbf{M_3}), \mathcal{B}(\mathbf{M_4}), \mathcal{B}(\mathbf{M_5})]) \tag{3}$$

where $[\cdot]$ is the channel-wise concatenation, and $\mathcal{B}$ denotes the bilinear interpolation. Finally, we calculate the binary cross-entropy (BCE) [16] loss between the overall saliency maps and the ground truth (GT) $\mathbf{Y}$ as below, and combine the losses

$$l_1 = \mathcal{L}_{\text{BCE}}(\mathbf{M_1}, \mathbf{Y}) \tag{4}$$

$$l_i = \mathcal{L}_{\text{BCE}}(\mathcal{B}(\mathbf{M_i}), \mathbf{Y})\ (i = 2, 3, 4, 5) \tag{5}$$

$$l_{\sum} = \mathcal{L}_{\text{BCE}}\left(\mathbf{M}_{\sum}, \mathbf{Y}\right) \tag{6}$$

$$L = \lambda_1 l_1 + \lambda_2 l_2 + \lambda_3 l_3 + \lambda_4 l_4 + \lambda_5 l_5 + \lambda_{\sum} l_{\sum} \tag{7}$$

in which λ_i $(i = 1, 2, 3, 4, 5)$ represents the weights corresponding to different loss functions. In this work, λ_i and $\lambda_{\sum}$ are set to 1 empirically.

B. Spatial–Channel Cross Transformer Block

Recently, successful architectures such as MLP-mixer [52] and Poolformer [53] have both considered the interaction between spatial and channel information in constructing context information. However, vanilla CCT focuses excessively on establishing channel information and overlooks the crucial role of spatial information in neighborhood modeling. To address this, we develop an SCTB as a spatial-channel blending unit to mix full-level encoded features. As shown in Fig. 3, given the ith level features $\mathbf{I_i} \in \mathbb{R}^{C_i \times h \times w}$, $(i = 1, 2, 3, 4)$, in which $h = (H/16)$, $w = (W/16)$. The procedure of SCTB can be defined as

$$\mathbf{J}_{\sum} = \text{LN}([\mathbf{I_1}, \mathbf{I_2}, \mathbf{I_3}, \mathbf{I_4}]) \tag{8}$$

$$\mathbf{J_i} = \text{LN}(\mathbf{I_i}) \tag{9}$$

$$\mathbf{P_i} = \text{SSCA}\left(\mathbf{J_1}, \mathbf{J_2}\mathbf{J_3}, \mathbf{J_4}, \mathbf{J}_{\sum}\right) + \mathbf{I_i} \tag{10}$$

$$\mathbf{O_i} = \text{CFN}_i(\mathbf{P_i}) \tag{11}$$

where LN denotes the layer normalization, $\mathbf{J_i} \in \mathbb{R}^{C_i \times h \times w}$, $(i = 1, 2, 3, 4)$ and the concatenated tokens $\mathbf{J}_{\sum} \in \mathbb{R}^{C_{\sum} \times h \times w}$ are the five inputs of SSCA, $\mathbf{P_i}$ represent the outputs of SSCA, and $\mathbf{O_i}$ stands for the outputs of SCTB. The SSCA: spatial-embedded single-head channel cross-attention and CFN: complementary feed-forward network are separately described below.

1) Spatial-Embedded Single-Head Channel Cross-Attention: In Fig. 3(a), given the five input tokens $\mathbf{J_i}$ and $\mathbf{J}_{\sum}$ for which LN is performed, the launching point of SSCA is to calculate the local-spatial channel similarity between single-level features and full-level concatenation features to establish global semantics. Therefore, our SSCA employs the four input tokens $\mathbf{J_i}$ as queries, one concatenated token $\mathbf{J}_{\sum}$ as key and value. This is accomplished by utilizing 1×1 convolutions to consolidate pixel-wise cross-channel context and then applying 3×3 depth-wise convolutions to capture local spatial context. Mathematically

$$\mathbf{Q_i} = W_{di}^{Q} W_{pi}^{Q} \mathbf{J_i},\ \mathbf{K} = W_d^K W_p^K \mathbf{J}_{\sum},\ \mathbf{V} = W_d^V W_p^V \mathbf{J}_{\sum} \tag{12}$$

where $W_{pi}^{(\cdot)} \in \mathbb{R}^{C_i \times 1 \times 1}$ and $W_p^{(\cdot)} \in \mathbb{R}^{C_{\Sigma} \times 1 \times 1}$ are the 1×1 point-wise convolution, $W_{di}^{(\cdot)} \in \mathbb{R}^{C_i \times 3 \times 3}$ and $W_d^{(\cdot)} \in \mathbb{R}^{C_{\Sigma} \times 3 \times 3}$ are the 3×3 depth-wise convolution. Next, we reshape $\mathbf{Q_i} \in \mathbb{R}^{C_i \times h \times w}$, $\mathbf{K} \in \mathbb{R}^{C_{\Sigma} \times h \times w}$, and $\mathbf{V} \in \mathbb{R}^{C_{\Sigma} \times h \times w}$ to $\mathbb{R}^{C_i \times hw}$, $\mathbb{R}^{C_{\Sigma} \times hw}$ and $\mathbb{R}^{C_{\Sigma} \times hw}$, separately. Our SSCA process is defined as

$$\mathbf{CA_i} = W_{\text{pi}}\ \text{CrossAtt}(\mathbf{Q_i}, \mathbf{KV}) \tag{13}$$

$$\text{CrossAtt}(\mathbf{Q_i}, \mathbf{K}, \mathbf{V}) = \mathbf{A_i V} = \text{Softmax}\left\{\mathcal{I}\left(\frac{\mathbf{Q_i}\ \mathbf{K^T}}{\lambda}\right)\right\}\mathbf{V} \tag{14}$$

where $\mathbf{CA_i} \in \mathbb{R}^{C_i \times h \times w}$ are the output of SSCA, $\mathbf{A_i} \in \mathbb{R}^{C_i \times C_{\Sigma}}$ represent different level covariance-based attention maps, $\mathcal{I}$ denotes the instance normalization operation [59], and λ is an optional temperature factor defined by $\lambda = (C_{\sum})^{1/2}$. Notably, we differ from the common channel cross-attention under two further aspects: our patches are without positional encoding, and we use a single head to learn the attention matrix. These strategies will be compared for their efficacy in detail in the ablation study in Section IV-E4.

2) Complementary Feed-Forward Network: As shown in Fig. 4(a), previous studies [30], [32], [41] always incorporate single-scale depth-wise convolutions into the standard feed-forward network (FFN) to enhance local focus. More recently, state-of-the-art (SOTA) MSFN [31] incorporates two paths with depth-wise convolution using different kernel sizes to enhance the multiscale representation. However, the above approaches are limited to an LSGC paradigm of feature representation. In fact, GSLC information [Fig. 4(b)] is equally important [60]. Hence, we design a CFN, which combines the advantages of both feature representations.

In Fig. 3(b), given an input tensor $\mathbf{X_i} \in \mathbb{R}^{C_i \times h \times w}$, CFN first models multiscale LSGC information. Specifically, after the layer normalization, CFN utilizes 1×1 convolution to increase the channel dimension in the ratio of η and splits the feature map equally into two branches. Subsequently, 3×3 and 5×5 depth-wise convolutions are employed to enhance the local spatial information. This is followed by channel concatenating the multiscale features and restoring them to their original dimensions. The above process can be defined as

$$\mathbf{X_{3\times3}}, \mathbf{X_{5\times5}} = \text{Chunk}\left(f_{1\times1}^{c}(LN(\mathbf{X_i}))\right) \tag{15}$$

$$\mathbf{X_{sc}} = f_{1\times1}^{c}\left[\delta\left(f_{3\times3}^{\text{dwc}}(\mathbf{X_{3\times3}})\right), \delta\left(f_{5\times5}^{\text{dwc}}(\mathbf{X_{5\times5}})\right)\right] \tag{16}$$

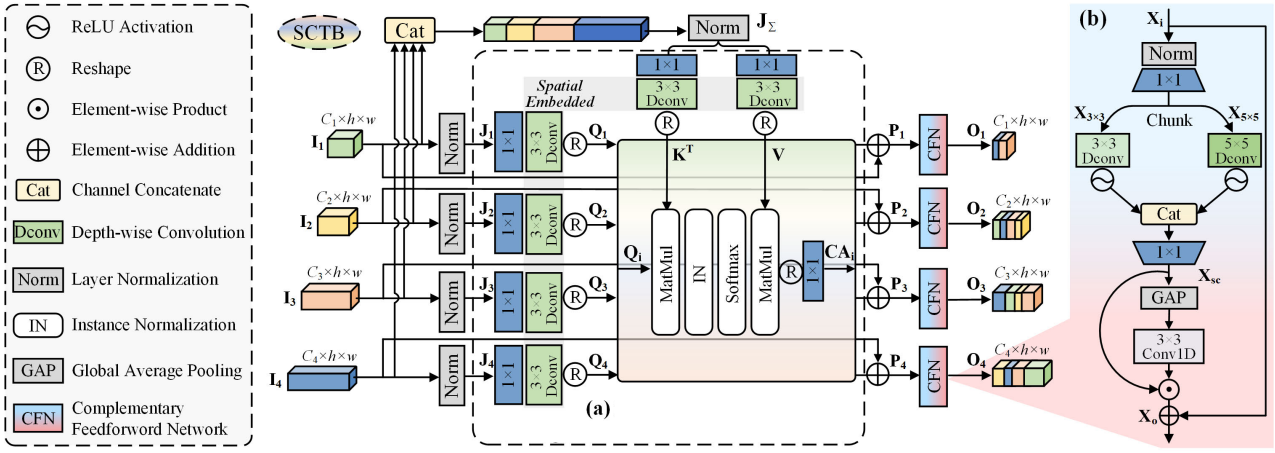


Fig. 3. Proposed SCTB, which consists of SSCA and CFN. (a) SSCA establishes image full-scale information association by means of different levels of semantic interaction. (b) CFN bridges the semantic gap between encoder and decoder through complementary feature enhancement. .

TABLE I

COMPARISONS WITH SOTA METHODS ON NUAA-SIRST, NUDT-SIRST AND IRSTD-1K IN $IoU(\%)$, $nIoU(\%)$, $F\text{-}Measure(\%)$, $Pd(\%)$, $Fa(10^{-6})$

Method	NUAA-SIRST [14]					NUDT-SIRST [15]					IRSTD-1K [40]				
	mIoU	nIoU	F-measure	Pd	Fa	mIoU	nIoU	F-measure	Pd	Fa	mIoU	nIoU	F-measure	Pd	Fa
Top-Hat [5]	7.143	18.27	14.63	79.84	1012	20.72	28.98	33.52	78.41	166.7	10.06	7.438	16.02	75.11	1432
Max-Median [54]	4.172	12.31	10.67	69.20	55.33	4.197	3.674	7.635	58.41	36.89	6.998	3.051	8.152	65.21	59.73
WSLCM [55]	1.158	6.835	4.812	77.95	5446	2.283	3.865	5.987	56.82	1309	3.452	0.678	2.125	72.44	6619
TLLCM [56]	1.029	4.099	4.995	79.09	5899	2.176	4.315	7.225	62.01	1608	3.311	0.784	2.186	77.39	6738
IPI [9]	25.67	50.17	43.65	84.63	16.67	17.76	15.42	26.94	74.49	41.23	27.92	20.46	35.68	81.37	16.18
PSTNN [57]	30.30	33.67	39.16	72.80	48.99	14.85	23.57	35.63	66.13	44.17	24.57	17.93	37.18	71.99	35.26
MSLSTIPT [1]	10.30	15.93	18.83	82.13	1131	8.342	10.06	18.26	47.40	888.1	11.43	5.932	12.23	79.03	1524
ACM [14]	68.93	69.18	80.87	91.63	15.23	61.12	64.40	75.87	93.12	55.22	59.23	57.03	74.38	93.27	65.28
ALCNet [18]	70.83	71.05	82.92	94.30	36.15	64.74	67.20	78.59	94.18	34.61	60.60	57.14	75.47	92.98	58.80
RDIAN [39]	68.72	75.39	81.46	93.54	43.29	76.28	79.14	86.54	95.77	34.56	56.45	59.72	72.14	88.55	26.63
ISTDU [22]	75.52	79.73	86.06	96.58	14.54	89.55	90.48	94.49	97.67	13.44	66.36	63.86	79.58	93.60	53.10
MTU-Net [17]	74.78	78.27	85.37	93.54	22.36	74.85	77.54	84.47	93.97	46.95	66.11	63.24	79.26	93.27	36.80
IAANet [58]	74.22	75.58	85.02	93.53	22.70	90.22	92.04	94.88	97.26	8.32	66.25	65.77	78.34	93.15	14.20
AGPCNet [19]	75.69	76.60	85.26	96.48	14.99	88.87	90.64	93.88	97.20	10.02	66.29	65.23	79.58	92.83	13.12
DNA-Net [15]	75.80	79.20	86.24	95.82	**8.78**	88.19	88.58	93.73	**98.83**	9.00	65.90	66.38	79.44	90.91	12.24
UIU-Net [16]	76.91	79.99	86.95	95.82	14.13	93.48	93.89	96.63	98.31	7.79	66.15	66.66	79.63	**93.98**	22.07
SCTransNet	**77.50**	**81.08**	**87.32**	**96.95**	13.92	**94.09**	**94.38**	**96.95**	98.62	**4.29**	**68.03**	**68.15**	**80.96**	93.27	**10.74**

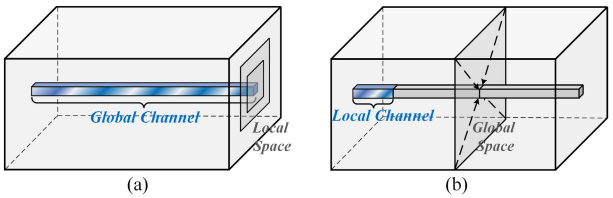


Fig. 4. Information enhancement from different perspectives. (a) LSGC paradigms. (b) GSLC paradigms. Our CFN integrates both of these information enhancement methods internally.

TABLE II

COMPREHENSIVE EVALUATION METRICS WITH COMPETITIVE ALGORITHMS

Model	Params (M)	Flops (G)	IoU	nIoU	F-measure
DNA-Net [15]	**4.697**	**14.26**	80.23	82.59	88.60
UIU-Net [16]	50.54	54.42	82.40	86.12	90.35
SCTransNet	11.19	20.24	**83.43**	**86.86**	**90.96**

where $f^{c}_{1\times1}$ denotes 1×1 convolution, $f^{\text{dwc}}_{3\times3}$ and $f^{\text{dwc}}_{5\times5}$ represent 3×3 and 5×5 depth-wise convolutions. Here, Chunk$(\cdot)$ denotes dividing the feature vector into two equal parts along the channel dimension.

Next, CFN constructs the GSLC information. Because of the varying resolution of the small target detection image inputs in the test stage, we first use the GAP of spatial dimensions to approximate the total spatial information of the features instead of using computationally intensive spatial MLPs to precisely compute the global spatial information [61]. We then employ a 1-D convolution with a kernel size of 3 to capture the local channel information of the spatially compressed feature as follows:

$$\mathbf{X_o} = f_3^{1D}(\text{GAP}_{2D}(\mathbf{X_{sc}})) \odot \mathbf{X_{sc}} + \mathbf{X_i} \tag{17}$$

where $\odot$ is the broadcasted Hadamard product. By incorporating complementary spatial and channel information, CFN enriches the representation of features in terms of the target's localization and the background's global continuity.

IV. Experiments and Analysis

A. Evaluation Metrics

We compare the proposed SCTransNet with the SOTA methods using several standard metrics.

1) *Intersection over union (IoU)*: IoU is a pixel-level evaluation metric defined as

$$\text{IoU} = \frac{A_i}{A_u} = \frac{\sum_{i=1}^{N} \text{TP}[i]}{\sum_{i=1}^{N} (T[i] + P[i] - \text{TP}[i])} \tag{18}$$

where A_i and A_u denote the size of the intersection region and union region, respectively. N is the number of samples, $TP[\cdot]$ denotes the number of true positive pixels, $T[\cdot]$ and $P[\cdot]$ represent the number of GT and predicted positive pixels, respectively.

2) *Normalized IoU (nIoU)*: nIoU is the normalized version of IoU [14], given as

$$\text{nIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}[i]}{T[i] + P[i] - \text{TP}[i]}. \tag{19}$$

3) *F-measure (F)*: It evaluates the miss detection and false alarms at pixel-level, given as

$$F = \frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}} \tag{20}$$

where Prec and Rec denote the precision rate and recall rate, respectively.

4) *Probability of detection (P_d)*: P_d is the ratio of correctly predicted targets N_{pred} and all targets N_{all}, given as

$$P_d = \frac{N_{\text{pred}}}{N_{\text{all}}}. \tag{21}$$

Following [15], if the deviation of target centroid is less than 3, we consider the target correctly predicted.

5) *False-alarm rate (F_a)*: F_a is the ratio of false predicted target pixels N_{false} and all the pixels in the image P_{all}, given as

$$F_a = \frac{N_{\text{false}}}{P_{\text{all}}}. \tag{22}$$

In addition to the fixed-threshold evaluation methods, we also utilize receiver operation characteristics (ROC) curves to comprehensively evaluate the models. ROC is used to describe the changing trends of P_d under varying F_a.

B. Experiment Settings

1) Datasets: In our experiments, we utilized three public datasets, namely NUAA-SIRST [14], NUDT-SIRST [15], and IRSTD-1K [40], which consist of 427, 1327, and 1000 images, respectively. We adopt the method used by [15] to partition the training and test sets of NUAA-SIRST and NUDT-SIRST, and [40] for splitting the IRSTD-1K. Hence, all splits are standard.

2) Implementation Details: We employ U-Net with four RBs as our detection backbone [17], the number of downsampling layers is 4, and the basic width is set to 32. The kernel size and stride size P for patch embedding are 16, the number of SCTB is 4, and the channel expansion factor η in CFN is 2.66. Our SCTransNet does not use any pre-trained weights for training, every image undergoes normalization and random cropping into 256×256 patches. To avoid over-fitting, we augment the training data through random flipping and rotation. We initialized the weights and bias of our model using the Kaiming initialization method [62]. The model is trained using the BCE loss function and optimized by the Adam optimizer with the initial learning rate of 0.001, and the learning rate is gradually decreased to 1×10^{-5} using the cosine annealing strategy. The batch size and epoch are set as 16 and 1000, respectively. Following [14], [15], [18], the fixed threshold to segment the salient map is set to 0.5. The proposed SCTransNet is implemented with PyTorch on a single Nvidia GeForce 3090 GPU, an Intel Core i7-12700KF CPU, and 32 GB of memory. The training process took approximately 24 h.

3) Baselines: To evaluate the performance of our method, we compare SCTransNet to the SOTA IRSTD methods, specifically seven well-established traditional methods (Top-Hat [5], Max-Median [54], WSLCM [55], TLLCM [56], IPI [9], PSTNN [57], MSLSTIPT [1]), and nine learning-based methods (ACM [14], ALCNet [18], RDIAN [39], ISTDU [22], IAANet [58], AGPCNet [19], DNA-Net [15], UIU-Net [16], and MTU-Net [17]) on the NUAA-SIRST, NUDT-SIRST, and IRSTD-1K datasets. To guarantee an equitable comparison, we retrained all the learning-based methods using the same training datasets as our SCTransNet, and following the original articles, adopted their fixed thresholds. Open-source implementations of most techniques can be found at https://github.com/XinyiYing/BasicIRSTD and https://github.com/xdFai/SCTransNet.

C. Quantitative Results

Quantitative results are shown in Table I. In general, the learning-based methods significantly outperform the conventional algorithms in terms of both target detection accuracy and contour prediction of targets. Meanwhile, our method outperforms all other algorithms. In the three metrics of *IoU*, *nIoU*, and *F-measure*, SCTransNet stands considerably ahead on all three public datasets. This indicates that our algorithm possesses a strong ability to retain target contours and can discern pixel-level information differences between the target and the background. We also note that even though SCTransNet does not obtain optimal P_d and F_a, e.g., DNA-Net's P_d is higher than ours by only 0.2 in the NUDT-SIRST, whereas our target detection false alarms are over twice as low as DNA-Net's. This demonstrates that our algorithm achieves a superior balance between false alarms and detection accuracy, as indicated by the remarkably high composite metric, *F-measure*. Next, we comprehensively compare the present algorithm with the most competitive deep learning methods, DNA-Net and UIU-Net. Table II gives the average metrics of the different algorithms on the three data, and we can

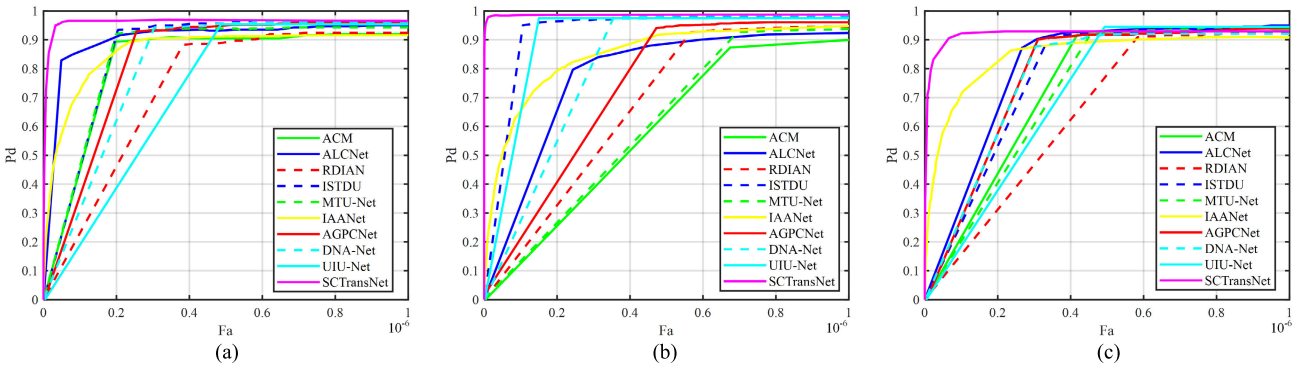


Fig. 5. ROC curves of different methods on the NUAA-SIRST, NUDT-SIRST, and IRSTD-1K dataset. Our SCTransNet can achieve the highest P_d at very low F_a. (a) NUAA-SIRST. (b) NUST-SIRST. (c) IRSTD-1K.

TABLE III
AUC WITH DIFFERENT THRESHOLDS OF THE SOTA METHODS ON THE NUAA-SIRST, NUDT-SIRST, AND IRSTD-1K DATASETS

Dataset	Index	ACM	ALCNet	RDIAN	ISTDU	MTU-Net	IAANet	AGPCNet	DNA-Net	UIU-Net	SCTransNet
NUAA-SIRST [14]	$AUC_{F_a=0.5}$	0.7223	0.8618	0.5461	0.7515	0.7457	0.8081	0.6953	0.6582	0.4854	**0.9539**
	$AUC_{F_a=1}$	0.8180	0.9025	0.7321	0.8579	0.8437	0.8614	0.8262	0.8098	0.7197	**0.9589**
NUDT-SIRST [15]	$AUC_{F_a=0.5}$	0.4392	0.6321	0.4630	0.8635	0.4640	0.7569	0.5038	0.6300	0.8275	**0.9853**
	$AUC_{F_a=1}$	0.5865	0.7716	0.6695	0.9211	0.6064	0.8463	0.7306	0.8072	0.9013	**0.9863**
IRSTD-1K [40]	$AUC_{F_a=0.5}$	0.5374	0.6606	0.4545	0.6014	0.5018	0.7862	0.6211	0.6162	0.4749	**0.9107**
	$AUC_{F_a=1}$	0.7366	0.8006	0.6480	0.7687	0.7198	0.8456	0.7752	0.7684	0.7099	**0.9200**

observe that SCTransNet has acceptable parameters at the highest performance and outperforms the powerful UIU-Net.

Fig. 5 displays the ROC curves of various competitive learning-based algorithms. It is evident that the ROC curve of SCTransNet outperforms all other algorithms. For instance, by appropriately selecting a segmentation threshold, SCTransNet achieves the highest detection accuracy while maintaining the lowest false alarms in the NUAA-SIRST and NUDT-SIRST datasets.

Table III presents the area under curve (AUC) of Fig. 5 in two different thresholds: $F_a = 0.5 \times 10^{-6}$ and $F_a = 1 \times 10^{-6}$. It can be seen that our method consistently achieves optimal detection performance across various false alarm rates. Meanwhile, while undergoing the same continuous threshold change, the curve of our method is more continuous and rounded compared to other methods. This observation suggests that SCTransNet showcases exceptional tunable adaptability.

D. Visual Results

The qualitative results of the seven representative algorithms in the NUAA-SIRST, NUDT-SIRST, and IRSTD-1K datasets are given in Figs. 6 and 7. Among them, conventional algorithms such as Top-Hat and TLLCM frequently yield a high number of false alarms and missed detections. Furthermore, even in cases where the target is detected, its contour is often unclear, hindering further accurate identification of the target type. In the learning-based algorithms, our method achieves precise target detection and effective contour segmentation. As illustrated in Fig. 6(2), our method successfully distinguishes between two closely located targets, whereas other deep learning methods tend to merge them into a single target.

TABLE IV
BASED ON U-NET, ABLATION STUDY OF THE RBS, DEEP SUPERVISION (DS), SSCA, CFN, AND CCA MODULE IN AVERAGE $IoU(\%)$, $nIoU(\%)$, AND $F\text{-}Measure(\%)$ ON NUAA-SIRST, NUDT-SIRST, AND IRSTD-1K

U-Net	+RBs	+DS	+SSCA	+CFN	+CCA	IoU	nIoU	F-measure
✓	✗	✗	✗	✗	✗	75.29	78.60	86.36
✓	✓	✗	✗	✗	✗	77.07	80.13	87.05
✓	✓	✓	✗	✗	✗	77.73	80.78	87.47
✓	✓	✓	✓	✗	✗	82.39	85.71	90.34
✓	✓	✓	✓	✓	✗	82.89	86.28	90.66
✓	✓	✓	✓	✓	✓	83.43	86.86	90.96

This suggests that our method discriminates each element in the image accurately. In Fig. 6(4), only our method accurately separates the shape of the unmanned aerial vehicle (UAV) from the mountain range. This is because our method not only learns the target's features but also constructs high-level semantic information about the backgrounds, thereby accurately capturing the overall continuity of the background. In Fig. 6(6), except for the present method and DNA-Net, the remaining methods produce false alarms on the stone in the grass. This can be attributed to their limitation in only constructing local contrast information and lack of establishing long-distance dependence on the image.

E. Ablation Study

In this section, we first employ two baselines to demonstrate the effectiveness of SCTransNet.

1) U-Net: We incrementally incorporate the RBs, deep supervised (DS), SSCA, CFN, and CCA into the baseline

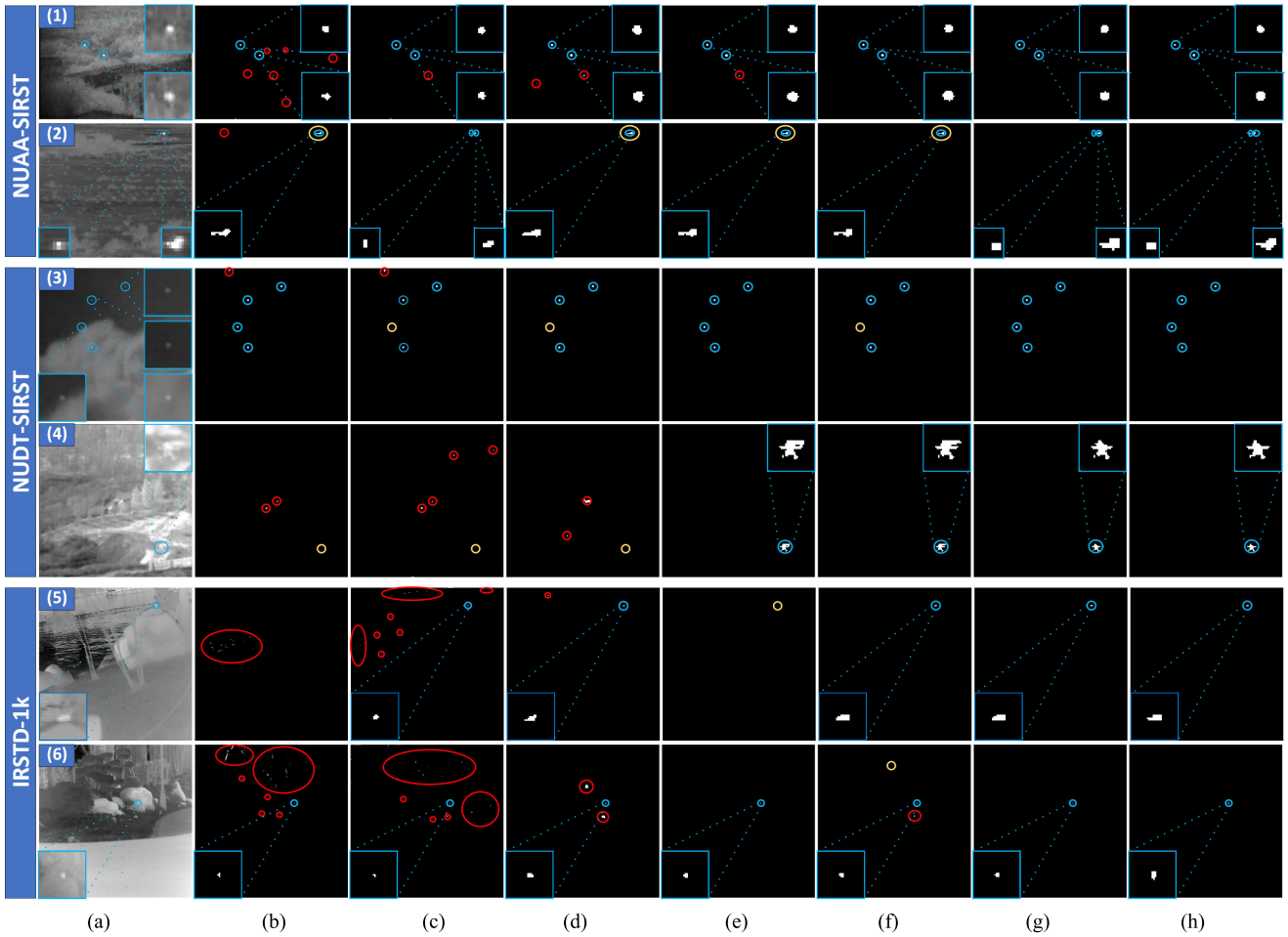


Fig. 6. Visual results obtained by different IRSTD methods on the NUAA-SIRST, NUDT-SIRST, and IRSTD-1K datasets. Circles in blue, yellow, and red represent correctly detected targets, miss detection, and false alarms, respectively. (a) Input. (b) Top-Hat. (c) TLLCM. (d) ACM. (e) DNA-Net. (f) UIU-Net. (g) SCTransNet. (h) GT.

TABLE V

BASED ON UCTRANSNET, ABLATION STUDY OF THE RBS, DS, SKS AND SCTB, REPORTING AVERAGE $IoU(\%)$, $nIoU(\%)$, AND $F\text{-}Measure(\%)$ ON NUAA-SIRST, NUDT-SIRST, AND IRSTD-1K. NOTE THAT, WE REPLACE THE CCT IN UCTRANSNET USING THE PROPOSED SCTB

UCTransNet	+RBs	+DS	+SKs	SCTB r/ CCT	IoU	nIoU	F-measure
✓	✗	✗	✗	✗	78.78	81.56	87.80
✓	✓	✗	✗	✗	79.95	82.97	88.45
✓	✓	✓	✗	✗	81.47	83.89	88.92
✓	✓	✓	✓	✗	82.03	84.98	89.54
✓	✓	✓	✓	✓	83.43	86.86	90.66

U-Net to validate the effectiveness of the above modules for IRSTD. The results are presented in Table IV. We observe that the algorithm's performance improves consistently with the inclusion of the aforementioned modules. In particular, the SSCA module significantly enhances the IoU, nIoU, and F-measure value of the algorithm by 4.66%, 4.93%, and 2.87%, respectively. This effectively demonstrates the effectiveness of the full-level information modeling of the IR small target.

2) UCTransNet: We incrementally incorporate the RBs, DS, and SKs, and use the proposed SCTB to replace CCT in the baseline UCTransNet to validate the effectiveness of these modules. As shown in Table V, these modules consistently enhance the algorithm's performance. Particularly, the proposed SCTB improves the IoU, nIoU, and F-measure value of the algorithm by 1.40%, 1.88%, and 1.12%, respectively, compared to the primitive CCT. This demonstrates the proposed SCTB can more effectively enhance the semantic difference between IR small targets and backgrounds than CCT.

Next, we will delve into a detailed discussion of the proposed SCTB, SSCA, and CFN, and compare the adopted CCA block with other feature fusion approaches implemented in IRSTD.

3) Spatial-Channel Cross Transformer Block: In the proposed SCTransNet, a primary idea is utilizing SCTB to mix and redistribute the output features of the full-stage encoders to predict contextual information about the small target and backgrounds. Since the network is encoded four times, the number of queries (Q) is set to 4, and both keys (K) and values (V) are formed by mapping the concatenated features (J) of the complete four-level features. In this section, we will discuss different levels of Q and the composition of J to illustrate the importance of full-level feature modeling.

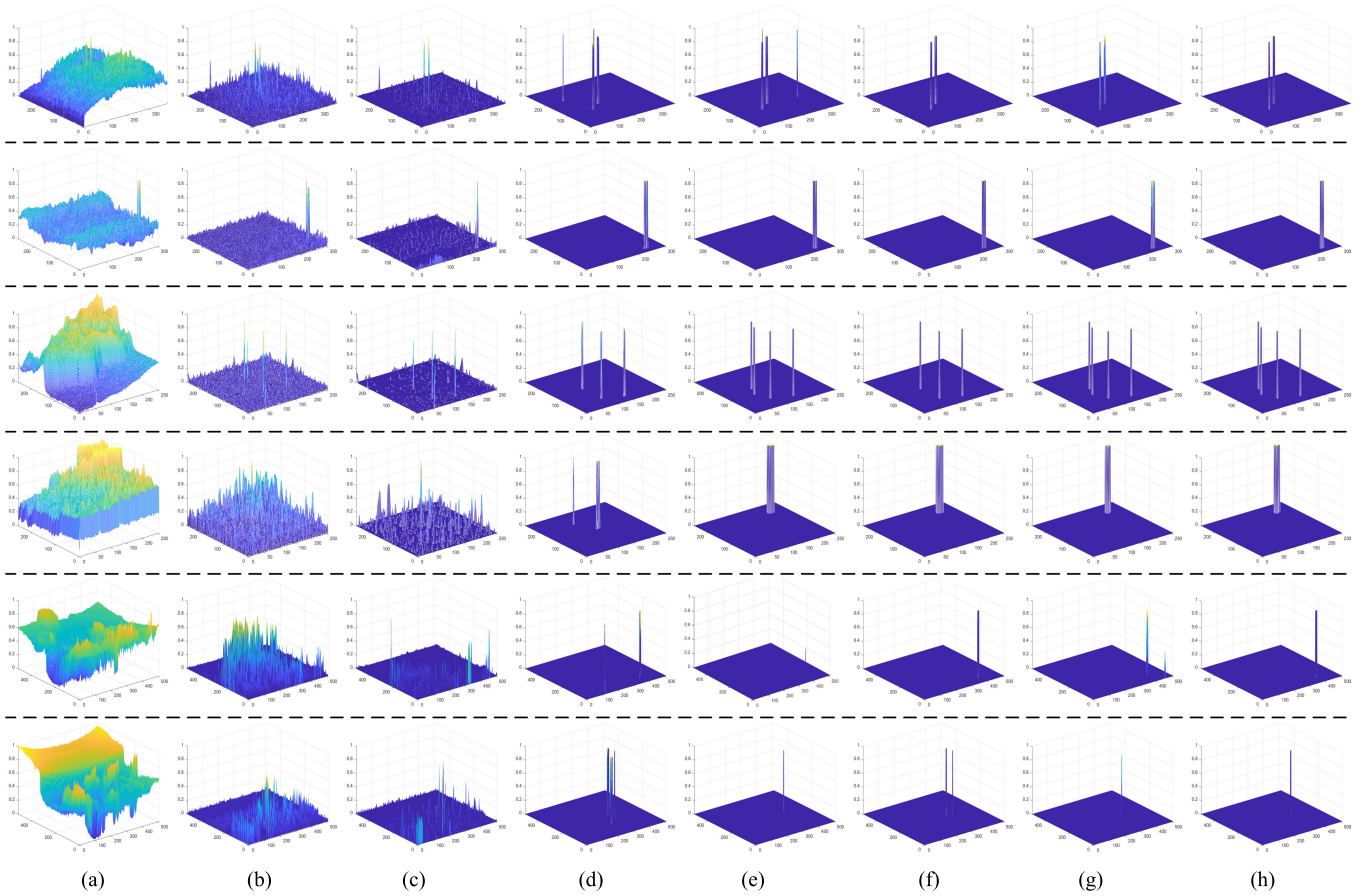


Fig. 7. 3-D visualization of salient maps of different methods on six test images. (a) Input. (b) Top-Hat. (c) TLLCM. (d) ACM. (e) DNA-Net. (f) UIU-Net. (g) SCTransNet. (h) GT.

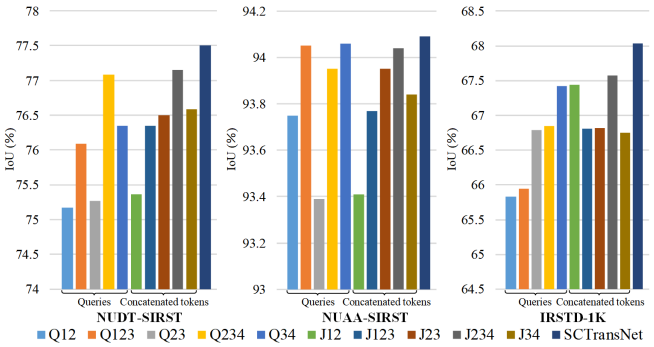


Fig. 8. Ablation on level of queries (Q) and composition of concatenated feature (J) on NUAA-SIRST, SUDT-SIRST, and IRSTD-1K.

Fig. 8 presents the ablation results for the level of Q and composition of J across three datasets. Note that when changing Q, J is composed of full-level features, and likewise, Q is the full-level feature input when varying J. The experimental results for Q indicate significant differences in the information learned by the neural network from different levels of features. Queries with higher and more comprehensive levels (Q123, Q234, Q34) encompass rich image semantics, thus achieving higher performance. The model performs best when fed with full-level Q inputs (SCTransNet), thus validating our motivation. Similarly, the experimental results for J suggest that selecting complete channel information allows queries to capture more accurate key features, thereby improving the performance of IRSTD.

4) Spatial-Embedded Single-Head Channel Cross-Attention: To demonstrate the efficacy of the proposed SSCA, we present multihead cross-attention [27] (MCA, a typical full-level information interaction structure in UCTransNet) and three network structure variants: SSCA with positional encoding (*SSCA w PE*), SSCA with multihead (*SSCA w MH*), and SSCA without spatial-embedding (*SSCA w/o SE*), respectively.

a) SSCA w PE: We incorporate positional encoding during the patch embedding stage. To accommodate test images of different sizes, we employ interpolation to scale the position-coding matrix, ensuring the proper functioning of the algorithm.

b) SSCA w MH: We use a typical multihead cross-attention mechanism to replace the single-head cross-attention mechanism in SSCA to verify the effectiveness of the single-head strategy for extracting limited features from the IR small targets.

c) SSCA w/o SE: To validate the effectiveness of local spatial information coding, we eliminate the depth-wise convolution in the QKV matrix generation process in SCTB.

As illustrated in Table VI, our SSCA has higher IoU, IoU, and F-measure values than the MCA and the variant *SSCA w PE* on three datasets. This suggests that SCTransNet can better

TABLE VI

IoU(%)/nIoU(%)/F-Measure(%) Values Achieved by Variants of SSCA and MCA on NUAA-SIRST, NUDT-SIRST, and IRSTD-1K

Model	Dataset		
	NUAA-SIRST	NUDT-SIRST	IRSTD-1K
MCA [27]	74.72/78.35/85.53	93.07/93.61/96.41	65.60/66.57/79.22
SSCA w PE	77.10/79.88/87.07	94.03/94.25/96.93	66.01/65.29/79.52
SSCA w MH	76.35/79.56/86.59	93.72/94.13/96.76	67.08/67.55/80.30
SSCA w/o SE	76.40/79.19/86.62	93.23/93.49/96.50	66.10/65.48/79.59
SSCA	**77.50/81.08/87.32**	**94.09/94.38/96.95**	**68.03/68.15/80.96**

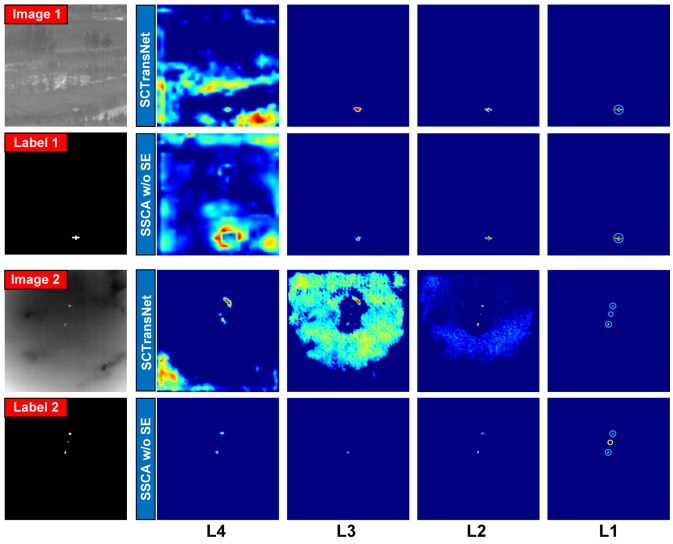


Fig. 9. Visualization map of SCTransNet and SSCA w/o SE. The feature maps from the deep layers of SCTransNet have an accurate representation of the localized region of the target from the background, and accurate segmentation results are obtained at the output layer.

perceive the information difference between small targets and complex backgrounds than MCA through comprehensive information interaction. It also illustrates that absolute positional encoding is not suitable for IRSTD tasks. This is due to the scaling of the position-embedding matrix in variable-size image inputs, which leads to inaccurate small-target position coding information, consequently affecting the prediction of target pixels.

Compared to our SSCA, *SSCA w MH* suffers decreases of 1.15%, 1.52%, and 0.73% in terms of IoU, IoU, and F-measure values on the SIRST-1K dataset. This is because the multihead strategy complicates the FM space of IR small targets, which is rather unfavorable for extracting information from targets with limited features. Therefore, in SCTransNet, we utilize the single-head attention for IRSTD.

Comparing SSCA and the variant *SSCA w/o SE*, we find that the local spatial embedding can significantly improve the performance of IRSTD in the three public datasets. Visualization maps displayed in Fig. 9 further illustrate the effectiveness of this strategy. This is due to the ability of local spatial embedding to capture both specific details of the target and potential spatial correlations in the background within the deep layers. As a result, this approach minimizes the instances of

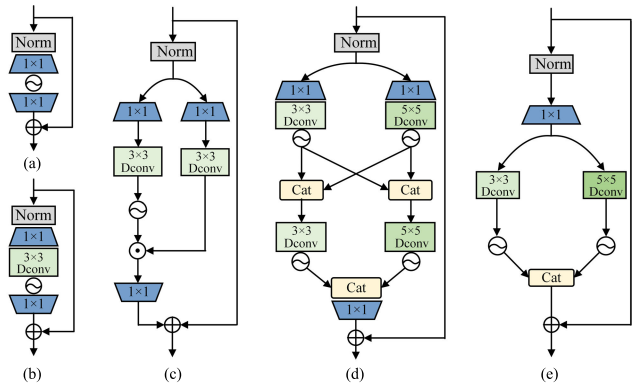


Fig. 10. Structure of representative FFNs and CFN w/o GSLC. (a) FFN. (b) LeFF. (c) GDFN. (d) MSFN. (e) CFN w/o GSLC.

TABLE VII

IoU(%)/nIoU(%) Values Achieved by the Representative FFNs and the Variants of CFN on NUAA-SIRST and NUDT-SIRST

Model	Params(M)	Flops(G)	Dataset	
			NUAA-SIRST	NUDT-SIRST
FFN [41]	**11.0292**	**20.1474**	76.87/80.08	93.58/93.85
LeFF [32]	11.1312	20.1944	76.49/80.21	93.92/94.07
GDFN [30]	10.1841	19.7210	75.48/79.32	93.40/93.64
MSFN [31]	11.7107	20.5026	77.35/79.89	93.88/94.24
CFN w/o GSLC	11.1905	20.2362	76.54/80.56	93.95/94.18
CFN	11.1905	20.2372	**77.50/81.08**	**94.09/94.38**

missed detections and improves the confidence of the detection process.

5) Complementary Feed-Forward Network: FFNs are used to strengthen the information correlation within features and introduce nonlinear radicalization to enrich the feature representation. In this section, we use five different FFN models based on SCTransNet to compare the proposed CFNs. As shown in Fig. 10, we used typical FFN [41] (ViT for image classification), LeFF [32] (Uformer for image restoration) embedded in localized space, GDFN [30] (Restormer for image restoration) based on gated convolution, MSFN [31] (Sparse transformer for image deraining) based on multiscale depth-wise convolution, the variant CFN without global spatial and local channel module (*CFN w/o GSLC*), respectively.

As shown in Table VII, LeFF exhibits a slight improvement in metrics over FFN, which indicates that the local spatial information aggregation employed in feed-forward neural networks is effective for IRSTD. Because gated convolution tends to consider IR small targets as noise and filters them out, this results in the GDFN having a low detection accuracy. We also find that MSFN outperforms all methods except our CFN, illustrating the superior ability of multiscale structures to interact with spatial information compared to single-scale structures. Finally, we observe that the performance of the variant *CFN w/o GSLC* is inferior to that of MSFN. However, when we incorporate the GSLC module, our CFN achieves optimal values of IoU and nIoU on the NUAA and NUDT datasets. Moreover, the network's parameters and computational complexity remain almost unchanged, which

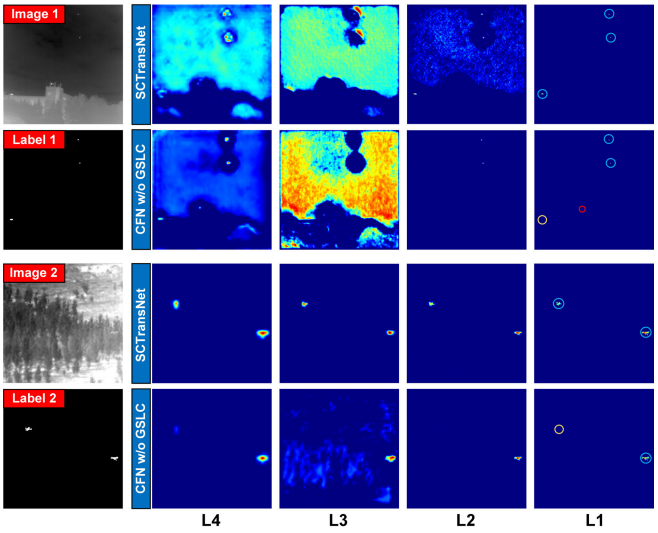


Fig. 11. Visualization map of SCTransNet and CFN w/o GSLC. The feature maps from the deep layer of CFN w/o GSLC mix targets in the background. It finally results in missed detection in the output layer.

TABLE VIII

IOU(%)/NIOU(%) VALUES ACHIEVED BY THE DIFFERENT CROSS-LAYER FEATURE FUSING MODULES ON NUAA-SIRST AND NUDT-SIRST

Model	Params(M)	Flops(G)	Dataset	
			NUAA-SIRST	NUDT-SIRST
C.ACM [14]	13.0627	30.9862	75.68/79.52	93.92/94.24
C.AGPC [19]	11.7581	22.9647	77.39/79.96	94.01/94.22
C.AFFPN [33]	11.7171	22.7291	76.12/79.33	93.53/93.69
SCTransNet	**11.1905**	**20.2372**	**77.50/81.08**	**94.09/94.38**

TABLE IX

HYPERPARAMETER STUDY OF THE RBS IN AVERAGE IOU(%), NIOU(%), F-MEASURE(%) ON NUAA-SIRST, NUDT-SIRST, AND IRSTD-1K

1	2	3	4	IoU	nIoU	F-measure	Params(M)	Flops(G)
✓	✗	✗	✗	82.29	85.77	90.26	20.0212	11.1462
✓	✗	✗	✗	82.33	85.89	90.31	20.0967	11.1484
✓	✓	✗	✗	82.49	86.11	90.40	20.1680	11.1569
✓	✓	✓	✗	82.95	86.27	90.68	20.2372	11.1905
✓	✓	✓	✓	83.43	86.86	90.96	20.2372	11.1905

demonstrates the validity and utility of the complementary mechanism proposed in this article for the IRSTD task. As illustrated in Fig. 11, with the help of the complementary mechanism, the network allows for more effective enhancement of infrared small targets and suppression of clutter in building and jungle backgrounds, leading to improved target detection accuracy.

6) Impact of CCA Block: As mentioned in Section II-A, cross-layer feature fusion can facilitate the preservation of enhanced target information. In this section, we utilize three cross-layer feature fusion structures, namely ACM [14], AGPC [19], and AFFPN [33], derived from different IRSTD methods, to replace the CCA module employed in SCTransNet. This substitution yields the variation structures, namely C.ACM, C.AGPC, and C.AFFPN, respectively. As shown in Table VIII, the results illustrate that our SCTransNet obtains the highest IoU and nIoU values on the NUAA and NUDT datasets with the lowest model parameters and computational complexity. This illustrates the effectiveness of the CCA we utilized.

TABLE X

HYPERPARAMETER STUDY OF THE NUMBER OF SCTBS, THE CHANNEL EXPANSION FACTOR OF CFN, AND THE BASIC WIDTH OF THE MODEL IN AVERAGE IOU(%), NIOU(%), F-MEASURE(%) ON NUAA-SIRST, NUDT-SIRST, AND IRSTD-1K

Hyper-param	IoU	nIoU	F-measure	Params(M)	Flops(G)
The number of SCTBs					
N = 1	82.33	85.86	90.28	17.7408	6.3295
N = 2	82.53	86.05	90.43	18.5729	7.9498
N = 3	82.97	86.46	90.58	19.4051	9.5702
N = 4	83.43	86.86	90.96	20.2372	11.1905
N = 5	83.40	86.84	90.95	21.0694	12.8108
N = 6	83.45	86.86	90.97	21.9015	14.4312
The channel expansion factor of CFNs					
η = 1.33	82.80	86.18	90.59	19.2457	9.2539
η = 2.00	82.75	86.32	90.56	19.7474	10.2338
η = 2.66	83.43	86.86	90.96	20.2372	11.1905
η = 3.00	83.24	86.69	90.84	20.4938	11.6917
η = 3.99	83.10	86.60	90.77	21.2306	13.1307
The basic width of the model					
W = 8	77.52	80.55	87.33	1.3321	0.7468
W = 16	81.02	84.50	89.51	5.1488	2.8609
W = 32	83.43	86.86	90.96	20.2372	11.1905
W = 48	82.95	86.48	90.60	45.2687	24.994

F. Core Hyperparameter Analysis

We utilize the depth of the RBs, the number of SCTBs, the channel expansion factor of CFNs, and the base width of the model to validate the hyperparameters of SCTransNet. As shown in Table IX, the numbers "0," "1," "2," and "3" indicate the embedding depth of the RBs. We observe that as the RB depth increases, there is a slight increase in both the number of parameters and flops, and the performance of IRSTD shows significant improvement. This improvement can be attributed to the residual connection facilitating gradient propagation and mitigating feature degradation. Therefore, our SCTransNet uses four RBs for information encoding. Table X illustrates the results of the hyperparameter study of the number of SCTBs, the channel expansion factor of CFNs, and the basic width of the model. It is evident that as the number of SCTB modules increases, the model's performance steadily improves, reaffirming the effectiveness of the SCTB model. We observe that while the performance with six SCTBs is slightly better than with four SCTBs, it incurs excessive computational complexity. When the channel expansion factor $\eta = 2.66$, the model can get the best performance. Additionally, we also noticed that setting the base width of the model W = 48 results in a slight degradation in performance compared with W = 32, which can be attributed to the excessive model parameters reducing the algorithm's generalization

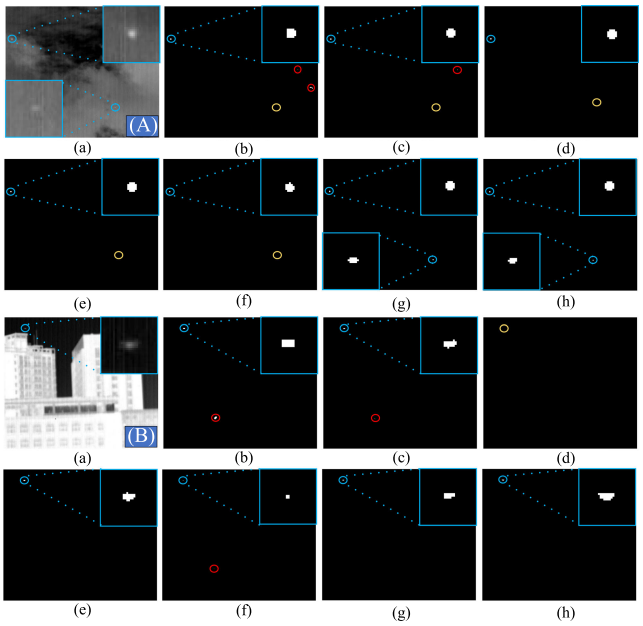


Fig. 12. Visual results obtained by different IRSTD methods on the real stripy scenes (A) and (B). Circles in blue, yellow, and red represent correctly detected targets, miss detections, and false alarms, respectively. (a) Stripy. (b) ACM. (c) RDIAN. (d) DNA-Net. (e) UIU-Net. (f) MTU-Net. (g) SCTransNet. (h) GT.

ability. Therefore, in our proposed SCTransNet, the number of SCTBs, the channel expansion factor of CFNs, and the base width of the model are set to 4, 2.66, and 32, respectively.

G. Robustness of SCTransNet

In an actual IR detection system, the nonuniform response of the focal plane array (FPN) can cause stripe noise in IR images [63]. This presents a challenge to the noise immunity and generalization ability of the IRSTD methods. Fig. 12 gives the visual effect of the IR image with real stripe noise on various detection methods. It is evident that the noise destroys the local neighborhood information of the targets. In Fig. 12(A), only our SCTransNet accurately detects two targets, while the other methods exhibit missed detections and false alarms. In Fig. 12(B), there is also a piece of blind element in the striped image, which interferes with the semantics understanding of the building. As a result, the ACM, RDIAN, and MTU-Net generate false alarms around the blind element. The ability to explicitly establish full-level contextual information about the target and the background is what makes our approach more robust.

V. Conclusion

In this article, we presented an SCTransNet for IR small target detection. Our SCTransNet utilizes SCTBs to establish associations between encoder and decoder features to predict the context difference of targets and backgrounds in deeper network layers. We introduced an SSCA module, which establishes the semantic relevance between targets and backgrounds by interacting local spatial features with global full-level channel information. We also devised a CFN, which employs a multiscale strategy and crosses spatial-channel information to enhance feature differences between the target and background, thereby facilitating effective mapping of IR images to the segmentation space. Our comprehensive evaluation of the method on three public datasets shows the effectiveness and superiority of the proposed technique.

References

[1] 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.

[2] P. Wu, H. Huang, H. Qian, S. Su, B. Sun, and Z. Zuo, "SRCANet: Stacked residual coordinate attention network for infrared ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5003614.

[3] P. Yan, R. Hou, X. Duan, C. Yue, X. Wang, and X. Cao, "STDMANet: Spatio-temporal differential multiscale attention network for small moving infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602516.

[4] X. Ying et al., "Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15528–15538.

[5] 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, Jun. 2010.

[6] J. F. Rivest and R. Fortin, "Detection of dim targets in digital infrared imagery by morphological image processing," *Opt. Eng.*, vol. 35, no. 7, pp. 1886–1893, Jul. 1996.

[7] C. P. 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. 2013.

[8] S. Kim and J. Lee, "Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track," *Pattern Recognit.*, vol. 45, no. 1, pp. 393–406, Jan. 2012.

[9] 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.

[10] H. Zhu, S. Liu, L. Deng, Y. Li, and F. Xiao, "Infrared small target detection via low-rank tensor completion with top-hat regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1004–1016, Feb. 2020.

[11] 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. (ICCV)*, Oct. 2019, pp. 8509–8518.

[12] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[13] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.

[14] 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.*, Oct. 2021, pp. 950–959.

[15] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2023.

[16] 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.

[17] T. Wu et al., "MTU-Net: Multilevel TransUnet for space-based infrared tiny ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[18] 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.

[19] 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, pp. 4250–4261, 2023.

[20] X. Tong et al., "MSAFFNet: A multiscale label-supervised attention feature fusion network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[21] M. Zhang, K. Yue, J. Zhang, Y. Li, and X. Gao, "Exploring feature compensation and cross-level correlation for infrared small target detection," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1857–1865.

[22] Q. Hou, L. Zhang, F. Tan, Y. Xi, H. Zheng, and N. Li, "ISTDU-Net: Infrared small-target detection U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[23] X. He, Q. Ling, Y. Zhang, Z. Lin, and S. Zhou, "Detecting dim small target in infrared images via subpixel sampling cuneate network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.* Cham, Switzerland: Springer, 2018, pp. 3–11.

[25] R. Kou, C. Wang, F. Huang, Y. Yu, Z. Peng, and Q. Fu, "LW-IRSTNet: Lightweight infrared small target segmentation network and application deployment," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[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 Represent.*, vol. 89, Nov. 2022, Art. no. 103684.

[27] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 2441–2449.

[28] Y. Li, Z. Cheng, C. Wang, J. Zhao, and L. Huang, "RCCT-ASPPNet: Dual-encoder remote image segmentation based on transformer and ASPP," *Remote Sens.*, vol. 15, no. 2, p. 379, Jan. 2023.

[29] Q. Luo, J. Su, C. Yang, W. Gui, O. Silvén, and L. Liu, "CAT-EDNet: Cross-attention transformer-based encoder–decoder network for salient defect detection of strip steel surface," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[30] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

[31] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5896–5905.

[32] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17683–17693.

[33] Z. Zuo et al., "AFFPN: Attention fusion feature pyramid network for small infrared target detection," *Remote Sens.*, vol. 14, no. 14, p. 3412, Jul. 2022.

[34] C. Yu et al., "Pay attention to local contrast learning networks for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[35] X. Tong, B. Sun, J. Wei, Z. Zuo, and S. Su, "EAAU-Net: Enhanced asymmetric attention U-Net for infrared small target detection," *Remote Sens.*, vol. 13, no. 16, p. 3200, Aug. 2021.

[36] S. Liu, P. Chen, and M. Woźniak, "Image enhancement-based detection with small infrared targets," *Remote Sens.*, vol. 14, no. 13, p. 3232, Jul. 2022.

[37] L. Huang, S. Dai, T. Huang, X. Huang, and H. Wang, "Infrared small target segmentation with multiscale feature representation," *Infr. Phys. Technol.*, vol. 116, Aug. 2021, Art. no. 103755.

[38] Y. Chen, L. Li, X. Liu, and X. Su, "A multi-task framework for infrared small target detection and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.

[39] H. Sun, J. Bai, F. Yang, and X. Bai, "Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[40] 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. (CVPR)*, Jun. 2022, pp. 877–886.

[41] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[42] M. Zhang, J. Zhang, R. Zhang, C. Wang, J. Guo, and X. Gao, "RKformer: Runge–Kutta transformer with random-connection attention for infrared small target detection," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 1730–1738.

[43] P. Pan, H. Wang, C. Wang, and C. Nie, "ABC: Attention with bilinear correlation for infrared small target detection," 2023, *arXiv:2303.10321*.

[44] F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, "Infrared small-dim target detection with transformer under complex backgrounds," 2021, *arXiv:2109.14379*.

[45] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[46] G. Chen, W. Wang, and S. Tan, "IRSTFormer: A hierarchical vision transformer for infrared small target detection," *Remote Sens.*, vol. 14, no. 14, p. 3258, Jul. 2022.

[47] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[48] M. Qi et al., "FTC-Net: Fusion of transformer and CNN features for infrared small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8613–8623, 2022.

[49] S. Meng, C. Zhang, Q. Shi, Z. Chen, W. Hu, and F. Lu, "A robust infrared small target detection method jointing multiple information and noise prediction: Algorithm and benchmark," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[50] C. Xu et al., "SCAD: A Siamese cross-attention discrimination network for bitemporal building change detection," *Remote Sens.*, vol. 14, no. 24, p. 6213, Dec. 2022.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[52] I. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.

[53] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10819–10829.

[54] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," *Proc. SPIE*, vol. 3809, pp. 74–83, Jul. 1999.

[55] J. Han et al., "Infrared small target detection based on the weighted strengthened local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1670–1674, Sep. 2021.

[56] J. Han, S. Moradi, I. Faramarzi, C. Liu, H. Zhang, and Q. Zhao, "A local contrast method for infrared small-target detection utilizing a tri-layer window," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1822–1826, Oct. 2020.

[57] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sens.*, vol. 11, no. 4, p. 382, Feb. 2019.

[58] 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.

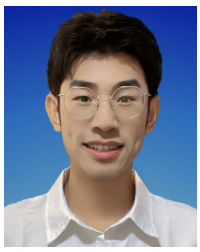
[59] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[60] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[61] H. Touvron et al., "ResMLP: Feedforward networks for image classification with data-efficient training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5314–5321, Apr. 2023.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2015, pp. 1026–1034.

[63] S. Yuan, H. Qin, X. Yan, N. Akhtar, S. Yang, and S. Yang, "ARCNet: An asymmetric residual wavelet column correction network for infrared image destriping," 2024, *arXiv:2401.15578*.



Shuai Yuan (Student Member, IEEE) received the B.S. degree from Xi'an Technological University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree with Xidian University, Xi'an.

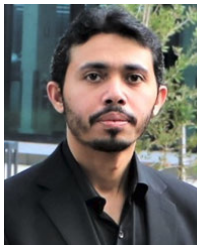
He is a Visiting Student with The University of Melbourne, Parkville, VIC, Australia, working closely with Dr. Naveed Akhtar. His research interests include infrared image understanding, remote sensing, and deep learning.

Hanlin Qin (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2010, respectively.

He is currently a Full Professor with the School of Optoelectronic Engineering, Xidian University. He authored or coauthored more than 100 scientific articles. His research interests include electro-optical cognition, advanced intelligent computing, and autonomous collaboration.

Xiang Yan (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2012 and 2018, respectively.

He was a Visiting Ph.D. Student with the School of Computer Science and Software Engineering, Perth, WA, Australia, from 2016 to 2018, working closely with Prof. Ajmal Mian. He is currently an Associate Professor with Xidian University. His research interests include image processing, computer vision, and deep learning.



Naveed Akhtar (Member, IEEE) received the master's degree from Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Germany, in 2012, and the Ph.D. degree in computer science from The University of Western Australia, Perth, WA, Australia, in 2016.

He is currently a Senior Lecturer with The University of Melbourne, Parkville, VIC, Australia.

Dr. Akhtar was a recipient of the Discovery Early Career Researcher Award from the Australian Research Council. He is a Universal Scientific Education and Research Network Laureate in Formal Sciences. He was a Finalist of the Western Australia's Early Career Scientist of the Year 2021. He serves as an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. He is an ACM Distinguished Speaker.

Ajmal Mian (Senior Member, IEEE) is currently a Professor of computer science with The University of Western Australia, Perth, WA, Australia. He has secured research funding from the ARC, NHMRC, DARPA, and the Australian Department of Defense. His research interests include computer vision, machine learning, remote sensing, and 3-D point cloud analysis.

Prof. Mian is a fellow of the International Association for Pattern Recognition. He was a recipient of three esteemed national fellowships from the Australian Research Council (ARC), including the recent Future Fellowship Award in 2022. He was a recipient of several awards, including the West Australian Early Career Scientist of the Year Award in 2012, the HBF Mid-Career Scientist of the Year Award in 2022, the Excellence in Research Supervision Award, the EH Thompson Award, the ASPIRE Professional Development Award, the Vice-Chancellors Mid-Career Research Award, the Outstanding Young Investigator Award, and the Australasian Distinguished Doctoral Dissertation Award. He has served as a Senior Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and *Pattern Recognition*.