FTC-Net: Fusion of Transformer and CNN Features for Infrared Small Target Detection

Meibin Qi, Liu Liu, Shuo Zhuang, Yimin Liu, Kunyuan Li, Yanfang Yang, and Xiaohong Li

***Abstract*—Single-frame infrared small target detection is still a challenging task due to the complex background and unobvious structural characteristics of small targets. Recently, convolutional neural networks (CNN) began to appear in the field of infrared small target detection and have been widely used for excellent performance. However, existing CNN-based methods mainly focus on local spatial features while ignoring the long-range contextual dependencies between small targets and backgrounds. To capture the global context-aware information, we propose fusion network architecture of transformer and CNN (FTC-Net), which consists of two branches. The CNN-based branch uses a U-Net with skip connections to obtain low-level local details of small targets. The transformer-based branch applies hierarchical self-attention mechanisms to learn long-range contextual dependencies. Specifically, the transformer branch can suppress background interferences and enhance target features. To obtain local and global feature representation, we design a feature fusion module to realize the feature concentration of two branches. We implement ablation and comparative experiments on a publicly accessed SIRST dataset. Experimental results show that the transformer-based branch is effective and suggest the superiority of the proposed FTC-Net compared with other state-of-the-art methods.**

***Index Terms*—Deep learning, feature fusion, hierarchical transformer, infrared small target detection.**

I. Introduction

INFRARED small target detection is increasingly applied in practical fields, including maritime surveillance [1], infrared warning, infrared guidance, infrared search, and tracking, and has made remarkable achievements. However, due to the lack of discriminative features such as color and texture, small size (less than 80 pixels in a 256 × 256 image [2]), long imaging distance, and low signal-to-noise ratio, infrared small targets are easy to be submerged by the noise in the complex and changeable background and cloud sea waves, making it difficult to detect. As a result, improving the detection rate of infrared small target detection tasks is still an inevitable demand for practical application.

To solve the above challenging task, many methods have been proposed, including filtering-based methods [3], [4], [5], vision-based methods [6], [7], [8], [9], and low-rank-based methods [10], [11], [12], [13]. The filtering-based method can suppress the uniform background. Nevertheless, the detection performance of small targets decreases when the background is complex, which means poor robustness. The vision-based method is mainly applicable to the scene where the target brightness is relatively large and different from the surrounding background. The low-rank-based method is time-consuming and has a high false alarm rate for infrared images with dark targets. The above methods rely on prior expert knowledge to extract handcraft features and are sensitive to varied scenarios.

Benefiting from the development of computer vision in many applications, the performance of the infrared small target detection method based on convolutional neural networks (CNN) is gradually improving. Liu et al. [14] proposed an end-to-end network based on multilayer perception to localize small targets. Fan et al. [15] enhanced infrared image contrast by applying a modified convolutional neural network and thus improved detection performance. Wang et al. [16] presented a conditional generative adversarial network that uses one discriminator and two generators to achieve a suitable balance of false alarm and miss detection. Specifically, segmentation-based methods for small target detection have begun to receive attention. Dai et al. [17] proposed an attentional local contrast network to capture long-range contextual interactions and applied a cross-layer fusion module to realize infrared small targets segmentation.

Most current image segmentation-based methods for detecting infrared small targets rely on the design of CNN architecture. However, the CNN-based methods ignore the long-range dependencies in infrared images. More specifically, convolutional networks tend to focus on the local information of an image, thus weakening the importance of the overall connection. Different from the standard CNN-based methods that process images pixel-by-pixel, a vision transformer (ViT) [18] treats an image as a series of patch tokens (i.e., smaller parts of the image consisting of multiple pixels). At each layer of the network, the ViT uses multihead self-attention to process patch tokens based on the relationship between each pair of tokens. As a result, the ViT can build a global representation of an entire image. Existing methods have achieved good performance, but the task of infrared small target segmentation is still worth exploring,

Manuscript received 9 June 2022; revised 14 August 2022 and 5 September 2022; accepted 15 September 2022. Date of publication 6 October 2022; date of current version 13 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 6227072073 and Grant 61771180, in part by the Hefei Municipal Natural Science Foundation under Grant 2021050, and in part by the Fundamental Research Funds for the Central Universities under Grant JZ2021HGQA0222. *(Corresponding author: Shuo Zhuang.)*

Meibin Qi, Liu Liu, Shuo Zhuang, Yimin Liu, Kunyuan Li, and Xiaohong Li are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China (e-mail: qimeibin@hfut.edu.cn; 2020171085@mail.hfut.edu.cn; shuozhuang@hfut.edu.cn; yiminliu@mail.hf-ut.edu.cn; kunyuan@mail.hfut.edu.cn; jsjlxh@hfut.edu.cn).

Yanfang Yang is with the School of Physics, Hefei University of Technology, Hefei 230601, China (e-mail: yyfatom@163.com).

Digital Object Identifier 10.1109/JSTARS.2022.3210707

Fig. 1. Comparison results of infrared small target detection. (a) Original infrared image containing a small target. (b) Predictions with CNN-based method. (c) Predictions with proposed FTC-Net. (d) Ground truth. The red circle area is detected incorrectly by the CNN-based method due to lack of long-range dependencies learning.

and further enhancements can be made in modeling long-range dependencies.

We made a comparison to illustrate that the long-range dependency is significant for the infrared small target segmentation task. To get the correct segmentation, the network needs to accurately distinguish whether the pixels of the image correspond to the background or the target. Due to the low signal-to-noise ratio of infrared images, background clutter can easily be mistaken for the target. Learning the long-range dependencies of targets and background interference in infrared images is critical, which helps to prevent the network from misclassifying a background pixel as a target pixel and reduces false positives. As shown in Fig. 1, due to the cloud background with the usual fuzzy and fractal structures on the cloud margins, the CNN-based method classifies a part of the background similar to the target (the red circle highlights the region) as the target. In contrast, our method avoids this error and predicts a more efficient result. This false positive situation is avoided because the proposed fusion network architecture of transformer and CNN (FTC-Net) learns the long-range dependencies between the target pixel region and the background.

To this end, we combine the sequence-model transformer with the CNN model to enhance the ability to capture long-range and large-range dependencies. Specifically, in the CNN-based branch, the modified U-Net with skip connections obtains the feature representation of local details to retain as many small targets as possible. The designed transformer branch is flexible and can be scaled to extract high-level global context information. Moreover, a feature fusion module (FFM) is designed to fuse local details with global contextual features. The experimental results show that our proposed method achieves the IoU and nIoU gains of nearly 2% and 4.2% on the SIRST dataset, respectively.

Our contributions can be summarized as follows:

- We design the FTC-Net to construct long-range dependencies and fully explore the global context between infrared small targets and background.
- We present an FFM to concatenate features extracted from the CNN branch and the transformer branch, which can obtain global context information while retaining location details as much as possible for detecting small targets.
- Experimental results on the SIRST dataset show the superior performance of the proposed FTC-Net, which is robust to clutter background, various target sizes and shapes.

The organization of this article is as follows: In Section II, we present the related work. In Section III, we illustrate the composition architecture of our FTC-Net. In Section IV, we show the experimental details and final results. In Section V, we draw the conclusion.

II. Related Work

A. Single-Frame Infrared Small Target Detection

In the field of computer vision, single-frame infrared small target detection is always an important topic and research hotspot. Typical traditional methods include filtering-based methods [19], vision-based methods [20], [21], [22], [23] and low-rank-based methods [24]. The filtering-based detection method highlights small targets by differencing the original image from the filtered background image. Alternatively, the frequency difference between the target, background, and clutter is used to filter out the background from the clutter by designing the corresponding filter in the frequency domain. The vison-based approach mainly utilizes the saliency map. According to the visual perceptual properties of the human eye, the presence of small targets leads to significant changes in local texture rather than global texture. As a result, the local difference or variation-based algorithms excel in different small target detection tasks. The low-rank basis has excellent stability performance for ordinary infrared images. However, for infrared images with dark targets, some strong clutter signals may be as sparse as the target signals, leading to higher false alarm rates.

Recently, CNN-based methods perform better than traditional methods. Liu et al. [14] proposed an end-to-end network based on multilayer perception to localize small targets. Wang et al. [25] proposed a feature extraction network with an attention mechanism while incorporating the YOLO [26] algorithm for target detection. Specifically, segmentation-based small target detection methods have begun to receive attention. To achieve a balance between miss detection and false alarm, Wang et al. [16] proposed a novel generative adversarial framework consisting of one discriminator and two generators. The segmentation results were obtained by computing the average outputs of two generators. Dai et al. [18] designed a segmentation-based infrared small target detection network, which proposed an asymmetric contextual modulation (ACM) framework for information exchange between high-dimensional features and low-dimensional features. In addition, Dai et al. [17] designed an attentional local contrast network, using a local attention module and a cross-layer fusion procedure to preserve local spatial features and enhance the segmentation performance of small targets.

Although the detection capacity of CNN-based approaches is gradually improving, the problem of effectively capturing long-range dependencies in infrared small target images remains a challenge.

B. ViT Transformers

Our work is motivated by the vision transformer (ViT) [18], which has achieved impressive performance in many tasks in the field of computer vision [27], [28], [29], [30], [31] and has

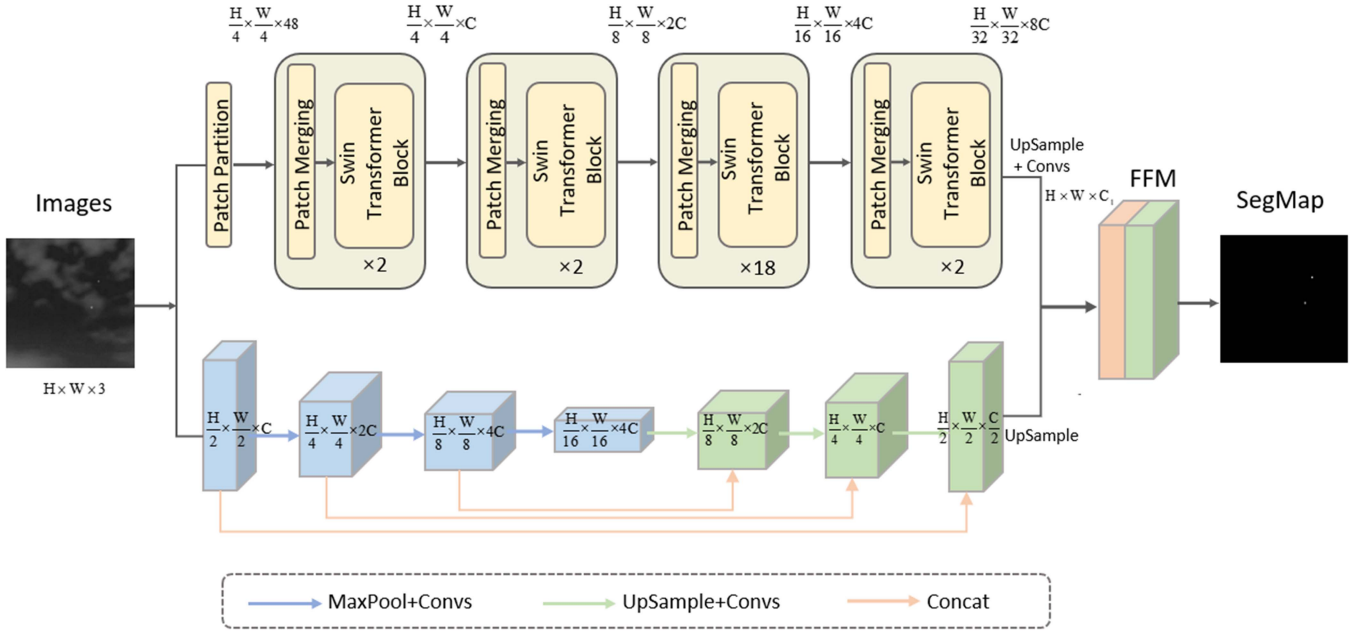

Fig. 2. Illustration of the proposed FTC-Net. W and H represent the width and height of the original image, "×2" and "×18" represent the number of Swin transformer blocks. In the CNN branch, the blue and green arrows represent the operation of max pooling and convolutions, upsampling, and convolutions.

been applied to target detection. In segmentation transformer [32], they integrated a transformer framework on the base of a fully convolutional network, and designed effective decoders to improve segmentation performance. The feature pyramid network in SOTR [33] can effectively distinguish low-level feature information, and the twin transformer can capture the association between remote contexts. In recent years, the framework combining CNN and transformer has been used in medical images. In two-dimensional (2D) medical image segmentation task [42], transformer and CNN are combined to form an enhanced encoder. In transfuse [43], the authors discovered the complementarity of transformer and CNN in image segmentation tasks. Recently, it is worth mentioning that an efficient hierarchical ViT architecture, called Swin transformer [34], whose representation was computed with shifted windows, achieved good performance on many vision tasks. In infrared small target detection, learning the long-range contextual dependencies of targets and background is critical [44], and transformer-based methods remain to be studied. While convolutional neural networks are still the main framework for all kinds of vision tasks, we must acknowledge that the potential of transformer-based architectures in this area cannot be ignored as well. In this work, we attempt to combine a hierarchical transformer and CNN model to detect infrared small targets.

III. Method

In this section, we introduce our FTC-Net in four parts. The model architecture of our method is shown in Fig. 2. We design an infrared small target detection network with two branches. Specifically, one branch uses a powerful hierarchical transformer to capture large-range dependencies, and the other branch uses a variant of U-Net to extract local details. Finally, a feature fusion module fuses the features extracted from the hierarchical transformer branch and the modified U-Net branch.

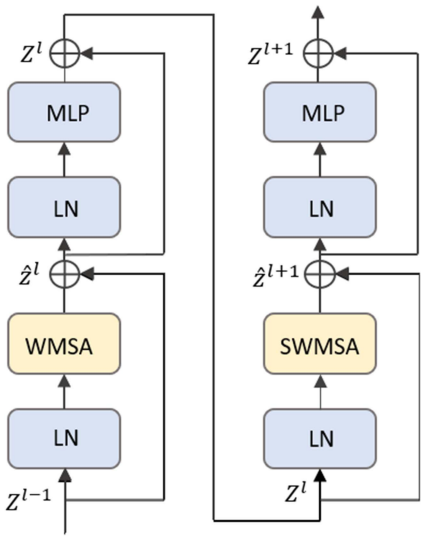

Fig. 3. Swin transformer block.

A. Architecture Overview

As shown in Fig. 2, for the transformer branch, we use a hierarchical construction with 4×, 8×, 16×, and 32× downsampling operations to obtain the image feature map with different sizes. To transform the inputs into sequence embedding's, the infrared input images are first to split into nonoverlapping patches with the size of 4 × 4, which is treated as a "token."

TABLE I
VALUE OF IOU, NIOU, AND PD ACHIEVED BY DIFFERENT METHODS ON SRIST DATASET. FOR IOU, NIOU, AND PD, LARGER VALUES INDICATE HIGHER PERFORMANCE

Method		IoU ($\times 10^{-2}$)	nIoU ($\times 10^{-2}$)	Pd ($\times 10^{-2}$)	AUC ($\times 10^{-2}$)
Local Contrast Based	LCM [35]	8.45	12.76	79.41	-
	FKRW [36]	25.96	30.23	85.79	-
	MPCM [37]	37.78	45.67	79.80	55.78
Local Rank Based	IPI [38]	39.66	56.71	75.21	56.49
	RIPT [12]	15.67	25.57	70.72	64.34
	NIPPS [40]	42.34	50.21	82.38	74.03
	PSTNN [41]	14.63	24.54	76.07	-
CNN Based	MDvsFA [16]	62.23	60.50	89.35	-
	ACM [42]	73.31	72.27	93.91	-
	ALC [17]	75.70	72.80	96.57	93.94
	FTC-Net	**77.72**	**77.02**	**99.05**	**96.58**

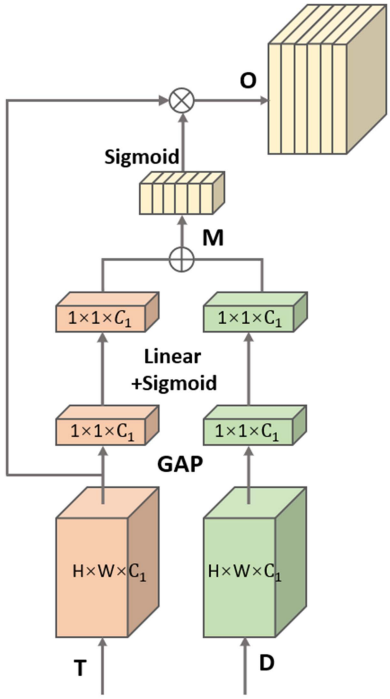

Fig. 4. FFM. T and D represent the output feature map of the transformer branch and the CNN branch. M represents the calculated attention mask, and O represents the output feature map after calibration.

Then the flatten operation is performed in the direction of channel. Therefore, the feature dimension of each patch is 48 ($4 \times 4 \times 3$). Through the linear embedding layer, the channel data of each feature map is linearly transformed from 48 to an arbitrary dimension C, which means the image shape is changed from $\left[\frac{H}{4}\frac{W}{4}, 48\right]$, to $\left[\frac{H}{4}\frac{W}{4}, C\right]$. By merging the layers, the feature dimension is halved, while the number of channels is doubled. Then the resolution remains the same when the feature transformation is performed with Swin transformer blocks. Therefore, the size of the feature maps output in the four stages are $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively.

For the CNN branch, we apply the U-Net with skip connections. Considering the contextual gap that exists from the encoder to the decoder, the feature representation with less local details may influence the final segmentation results. The skip connections in the U-Net help to recover fine-grained details in the segmentation task.

To better fuse the output features of the two branches and reduce the semantic gap, we design an FFM after the CNN branch and the transformer branch in the FTC-Net framework.

B. Swin Transformer Block

Multihead self-attention (MSA) is used to obtain multiple information from multiple inputs. Each attention module focuses on a different aspect of the feature map and finally gets a combined result to obtain relevant information on different subspaces. The Swin transformer block replaced the traditional MSA module with shifted windows. Fig. 3 shows the specific structure of the Swin transformer block, which consists of a LayerNorm (LN) layer, a shifted window-based multihead self-attention module (SW-MSA), a residual connection, and a multilayer perception (MLP) with the nonlinearity GELU function. The MLP Head contains two fully connected layers. The idea of layer normalization is similar to batch normalization. Compared with batch normalization that takes mini-batch size samples per neuron, LN normalizes the input of all neurons at a certain layer in the deep network. Layer normalization calculates the mean μ^l and standard deviation σ^l of all neurons separately for each sample as follows:

$$\mu^l = \frac{1}{H}\sum_{i=1}^{H} a_i^l \tag{1}$$

$$\sigma^l = \sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(a_i^l - \mu^l\right)^2} \tag{2}$$

where l, a_i^l, H represents the l^{th} hidden layer of the forward neural network, the input vector (the weighted vector of the

TABLE II
ABLATION EXPERIMENTS ON STRST DATASETS. "CNN," "TRANSFORMER," AND "FTC-NET" DENOTES THE CNN BRANCH, TRANSFORMER BRANCH AND PROPOSED METHOD WITH FEATURE FUSION MODULE

Method	IoU ($\times 10^{-2}$)	nIoU ($\times 10^{-2}$)	Pd ($\times 10^{-2}$)	Running time(s)
CNN (baseline)	74.09	75.66	96.33	0.238
Transformer	76.69	76.79	97.25	0.279
FTC-Net (CNN+Transformer+FFM)	**77.72**	**77.02**	**99.05**	0.313

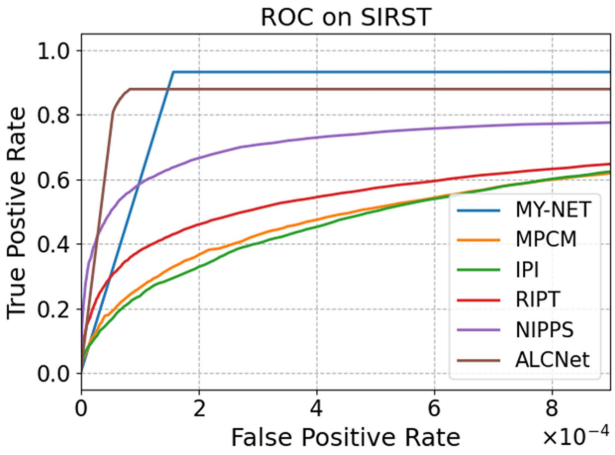

Fig. 5. ROC of different methods.

output of the front layer network), and the number of hidden units, respectively.

In the two successive transformer Blocks, the regular window-based multihead self-attention (WMSA) module and the SWMSA module appear alternately. WMSA divides the input image into nonoverlapping windows, and then performs self-attention calculations within different windows. Unlike a regular window, the shifted window is offset and remerged on the image. The advantage of this design is to ensure the window information interaction and reduce the computational complexity. To strengthen connections across windows, the partitioning mechanism for the regular window and shifted window alternate in the hierarchical Swin transformer block. The calculation formula is as follows:

$$\hat{z}^l = \text{WMSA}\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \tag{3}$$

$$z^l = MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l \tag{4}$$

$$\hat{z}^{l+1} = SWMSA\left(LN\left(z^l\right)\right) + z^l \tag{5}$$

$$z^{l+1} = MLP\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{6}$$

where $\hat{z}^l$ and z^l represent the outputs of the WMSA module and the multilayer perception module of the l^{th} block, respectively.

In addition, we add relative position bias for each head during the calculation of similarity:

$$Attention\ (Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{7}$$

where Q, K, V $\in R^{M^2 \times d}$ denote the query, key and value matrices; d is the dimension of the query or key, and M^2 is the number of patches in each window, respectively. We calculate the offset between the absolute position of each pixel and other positions, obtain the relative position index, and find the bias matrix $\hat{B}$ that can be learned according to the index value. Bias matrix B is added directly to the attention matrix, and the corresponding values are derived from the bias matrix $\hat{B} \in R^{(2M-1)\times(2M+1)}$.

C. Feature Fusion Module

To address the feature and semantic inconsistencies between the hierarchical transformer and CNN decoder outputs, we use an FFM to eliminate discrepancies.

As shown in Fig. 4, the FFM has two inputs, which are the output T $\in R^{C\times H\times W}$ of the hierarchical transformer and the output feature map D $\in R^{C\times H\times W}$ of the variant U-Net. For an input feature map X $\in X^{C\times H\times W}$, spatial squeeze generates a vector G(X) $\in G(X)^{C\times H\times W}$ through the global average pooling layer. Each element in G is calculated as follows:

$$G_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X^k(i, j) \tag{8}$$

This formula adds up all the elements in X and then divides them by a total dimension. Its actual meaning is to compress all the spatial information on the k^{th} channel into a value, and G(X) is a tensor obtained after compressing spatial information one by one on all channels. We combine the obtained tensor with the spatial information to calculate the attention mask:

$$M = L_1 \cdot G(T) + L_2 \cdot G(D) \tag{9}$$

where $L_1 \in R^{C\times C}$ and $L_2 \in R^{C\times C}$ are the weights of two linear layers. Then the activation value $\sigma(M)$ is obtained through the sigmoid layer so that it is between [0, 1]. The activation value allows us to distinguish the importance of channels, which is then multiplied with input T to obtain a feature map that has been calibrated with information:

$$O = \sigma(M) \cdot T \tag{10}$$

D. Loss Function

Two loss functions are applied in the proposed model, including the binary cross-entropy (BCE) loss and the Dice loss.

The segmentation-based infrared target detection method can be regarded as a binary classification task (target versus background). As a result, the BCE loss function is applied and

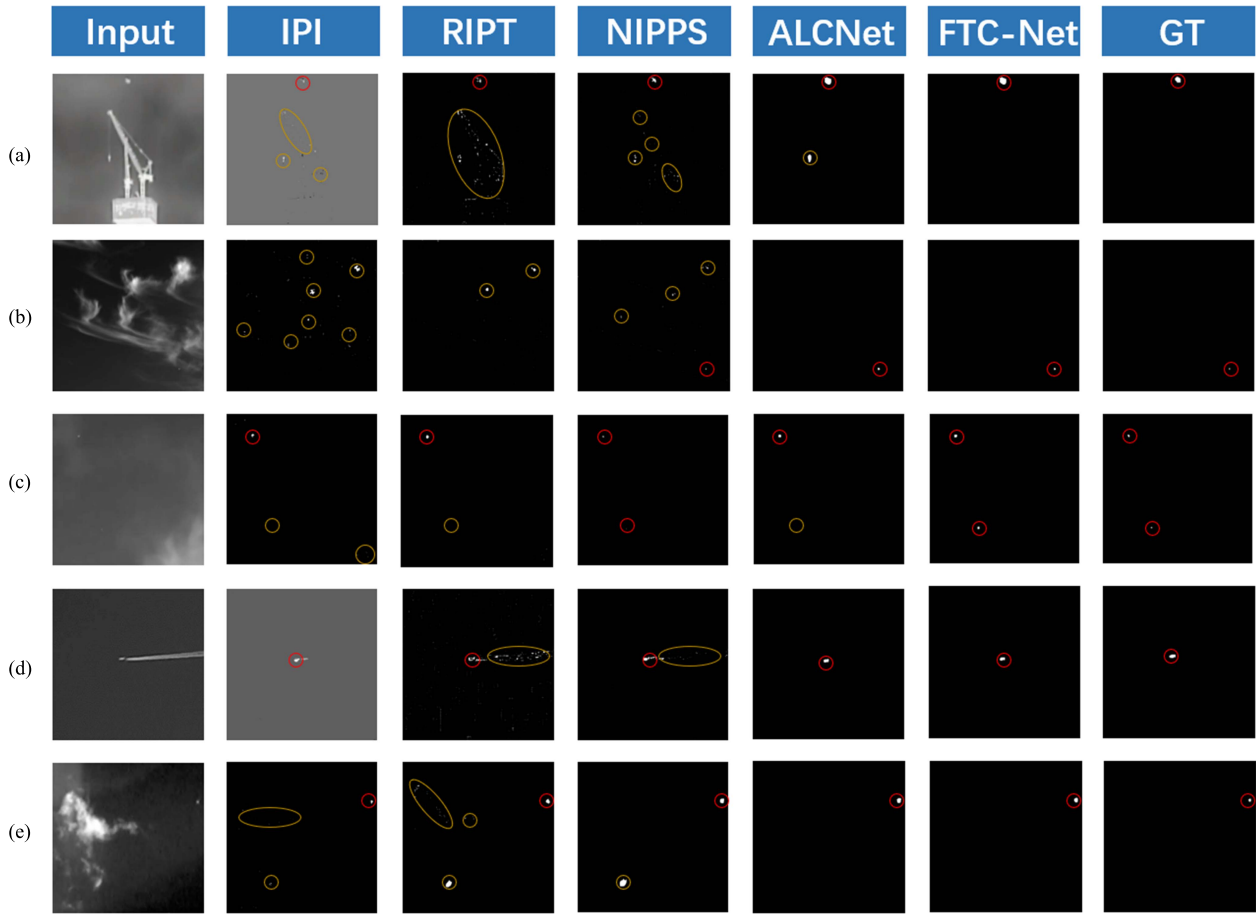

Fig. 6. Qualitative results obtained by different detection methods. For better visualization, correct detections (true positives) are highlighted with red circles, false alarms, and missed detections are highlighted with yellow circles. Our FTC-Net can achieve accurate target positioning and low FPR of output.

calculated as follows:

$$L_{BCE}\ (x, y) = \ L\ = \{l_1, \ldots, l_n, \ldots, l_N\}^T \tag{11}$$

$$l_n = \ -w_n \left[y_n \cdot \log x_n + (1 - y_n) \cdot \log (1 - x_n)\right] \tag{12}$$

where N is the sample number.

Dice loss is a common evaluation metric in medical image segmentation tasks [45], which is similar to an infrared small target detection application. It comes from a similarity function in statistics, which is proposed to measure the similarity of two samples and gives a value of 0 or 1. The formula of Dice loss is as follows:

$$L_{Dice} = \ 1 - 2\left|X \cap Y\right| / \left(|X| + |Y|\right) \tag{13}$$

where X and Y represent the predicted pixels and the ground truth pixels, respectively.

The overall optimization objective of proposed model is:

$$L\ = \ \alpha L_{BCE} + \beta L_{Dice} \tag{14}$$

where α and β are the balance coefficients.

IV. Experiments

A. Dataset

We evaluate our method using the SIRST dataset proposed by Dai et al. [41]. This open dataset is constructed with high-quality images and labels. It contains 427 single-frame infrared images, including short, medium, and 950 nm wavelengths, which are roughly divided into 70% for training, 10% for validation, and 20% for testing. In this dataset, the targets mainly appear in complex backgrounds such as sky, ocean, and city. It is easy to find that these small IR targets are relatively faint pixels, and some are distributed in the background without distinguishable characteristics. Considering this challenging detection task, we perform data enhancement by filtering, cropping, and horizontal and vertical flip methods to increase sample size and avoid overfitting in the training process.

B. Evaluation Metrics

Two evaluation metrics are explored for testing infrared small target detection performance. On the one hand, we use pixel-level evaluation metrics such as IoU and nIoU, which mainly focus on the shape of target segmentation.

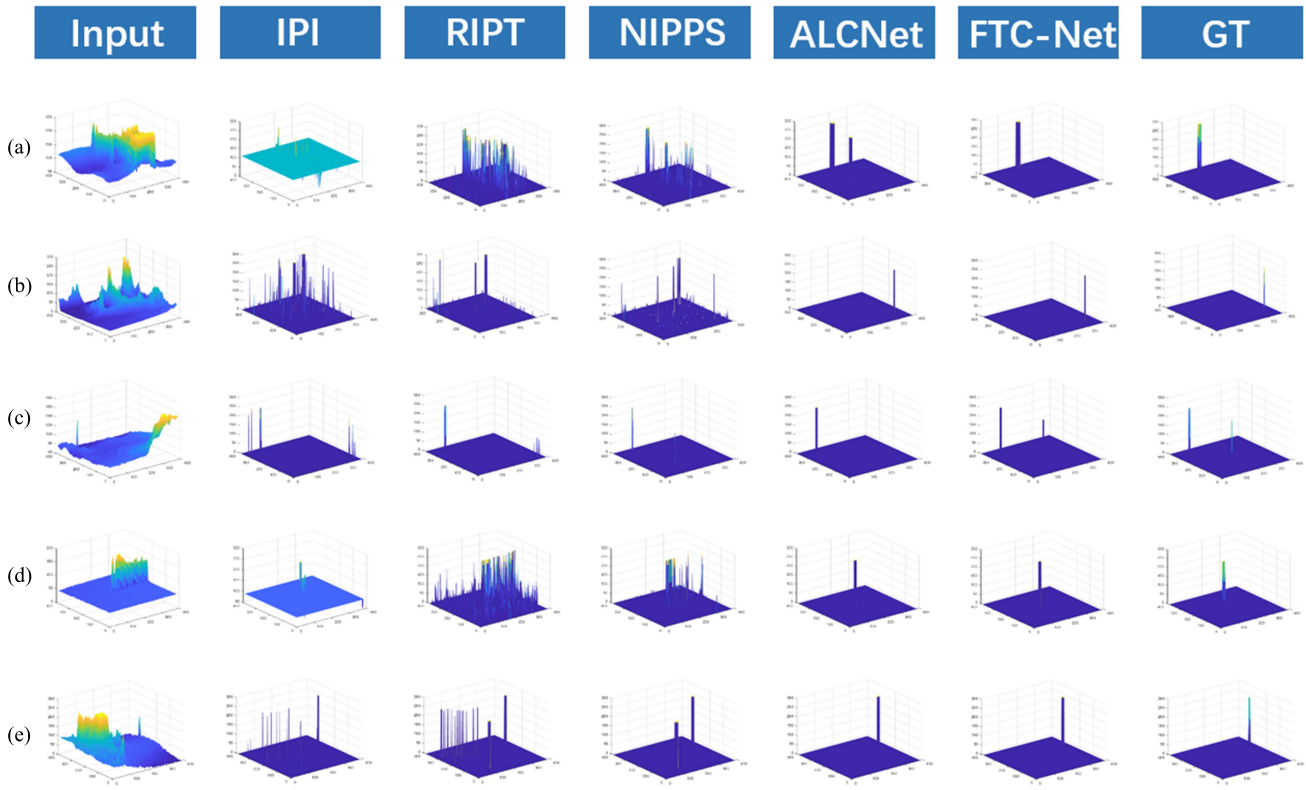

Fig. 7. 3D visualization results of different methods on five test images.

On the other hand, we use the target-level evaluation metric probability of detection (Pd) to assess the localization ability of the proposed model for infrared small targets. Generally, infrared small targets lack shape and texture information, a mispredicted pixel will result in a Pd decrease of 11.1% when evaluating a 3×3 small target.

1) IoU: Intersection over Union (IoU) is one of the most widely used metrics in the field of segmentation. Specifically, IoU represents the area of overlap between the labeled and predicted pixels divided by the union area between the predicted and labeled pixels.

$$IoU = \frac{A_{inter}}{A_{All}} \tag{15}$$

where A_{inter} and A_{All} represent the interaction areas and union areas, respectively.

2) nIoU: The average value of IoU of all targets in the images at a certain judgment threshold is used as the nIoU metric. nIoU is defined as follows:

$$nIoU = \frac{1}{N} \sum_{i}^{N} \frac{TP[i]}{T[i] + P[i] + TP[i]} \tag{16}$$

where N, TP, T, and P denote the number of total samples, true positive, true, and positive samples, respectively.

3) Probability of Detection: Pd is a widely used target-level metric. It represents the probability of successful detection of the target.

$$Pd = \frac{P_{correct}}{P_{All}} \tag{17}$$

where $P_{correct}$ and P_{All} are the number of correctly detected targets and all targets, respectively.

4) ROC: The horizontal and vertical axis of the receiver operation characteristics curve (ROC) are false positive rate (FPR) and true positive rate (TPR), respectively. FPR represents the probability of being misclassified as positive among all actually negative samples. TPR represents the probability of being classified as positive among all positive samples. ROC is a measure under a sliding threshold, which can effectively reflect the overall target detection performance.

C. Implementation Details

The proposed FTC-Net consists of a CNN-based segmentation branch and a hierarchical transformer branch. In the transformer branch, the number of down-sampling layers is set to 4. In the four Swin transformer blocks, the channel C ,C_1 is set to 128, 64. The number of MSA head is set to 4, 8, 16, and 32, respectively. As shown in Fig. 2, the number of the Swin transformer block is set to 2, 2, 18, and 2, respectively. The input image datasets with the different resolutions are first resized into a fixed size of 384×384. All models are implemented using PyTorch library on a computer with four Nvidia GeForce 1080Ti GPUS and are trained from scratch. We set the batch size to 8

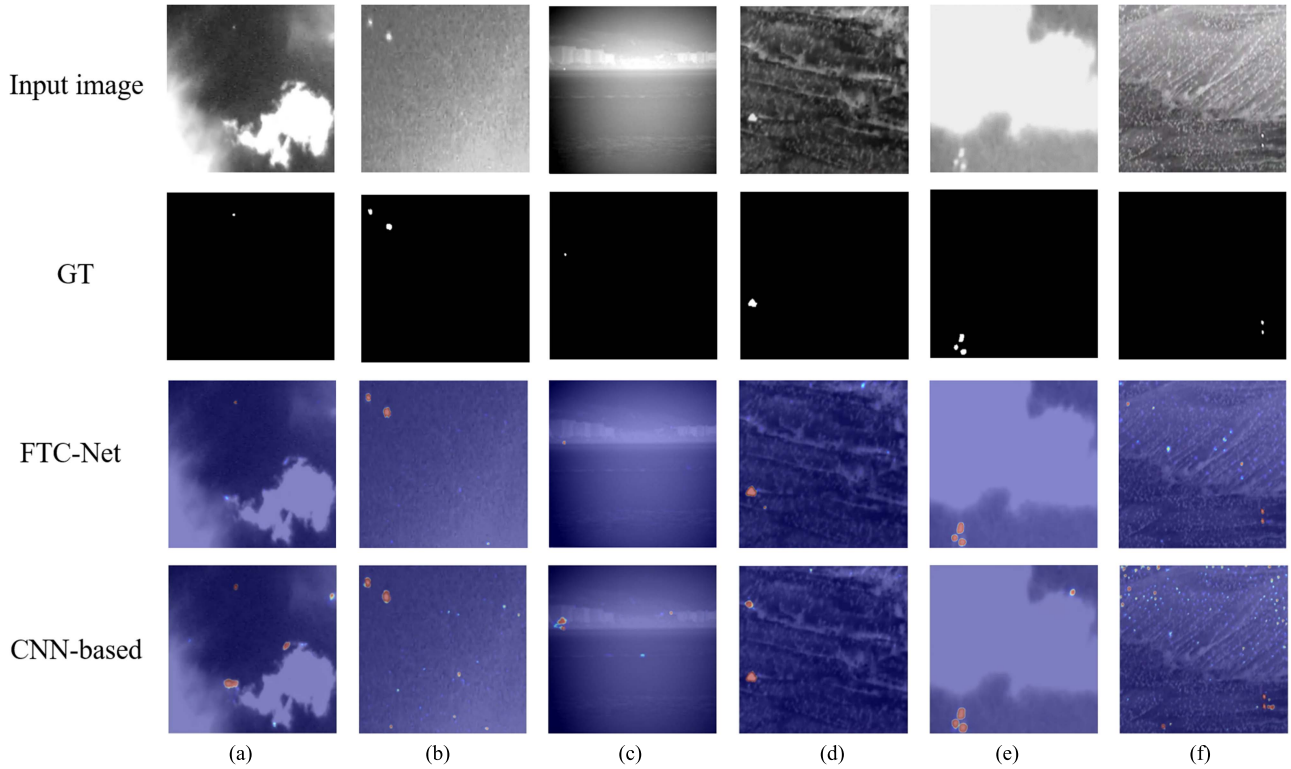

Fig. 8. Grad-CAM visualization map of FTC-Net (column 2) and CNN-based network (column 3).

and use the Adam optimizer with an initial learning rate of 0.001 to train the designed model.

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

To validate the superiority of the proposed FTC-Net, we compare it to several state-of-the-art methods, including several traditional methods: LCM [35], FKRW [36], MPCM [37], IPI [11], RIPT [12], NIPPS [39], PSTNN [40], and three CNN-based methods: MDvsFA-cGAN [16], ACM [41], and ALCNet [17].

1) Quantitative Results: We used the same dataset, publicly released code and settings in our experiments to ensure the authenticity of the comparison. Experimental results can be observed in Table I.

The prediction results using the above approaches are evaluated with IoU, nIoU, Pd, and ROC. Compared with traditional methods, the CNN-based detection framework has a significant improvement in all evaluation metrics. This is because the SIRST dataset has clutter in the background and contains challenging images with different target shapes and target sizes. The CNN-based methods are robust to changing backgrounds. In contrast, the local contrast and rank-based methods are generally scene-specific and can only suppress uniform backgrounds to a certain extent. In addition, traditional methods mainly focus on overall target positioning rather than fined shape matching, which gains relatively poor performance.

Compared with the CNN-based methods, our FTC-Net has achieved obvious improvement in detection performance. The designed network combines the advantages of both CNN and transformer, while taking into account the deep and shallow layers of the network as well as long-range dependencies. Therefore, the network can ensure accurate localization and precise segmentation of small infrared targets. As shown in Table I, compared with the state-of-the-art ALC method, our proposed FTC-Net has consistent improvements with IoU, nIoU, and Pd gains of approximately 2.02%, 4.22%, and 2.48%, respectively.

We further evaluate our FTC-Net and other existing methods using the ROC metric. The area under the ROC curve (AUC) is a measure of the performance of the classification model, and a larger value means better performance. The results are shown in Fig. 5 and Table I. It can be seen that the proposed FTC-Net has the best results, indicating the effectiveness of our network. The ROC of conventional methods is under the CNN-based methods, which means relatively poor performance. Compared to the ALCNet method with the attentional local contrast module, the presented FTC-Net has a larger area under the ROC curve with a value of 0.9658. The results further demonstrate the effectiveness and robustness of our method in infrared small target detection.

2) Qualitative Results: To visually understand the detection performance, qualitative results of five representative methods on the SIRST dataset are shown in Figs. 6 and 7. As can be seen, the ALC model and our FTC-Net are significantly superior to the traditional methods, which obtain more accurate segmentation results. As shown in Fig. 6(c) and 6(d), the traditional method performs well on images where infrared targets are clearly distinguished from the background. However, it is easy to miss the detection when the targets are similar to the background.

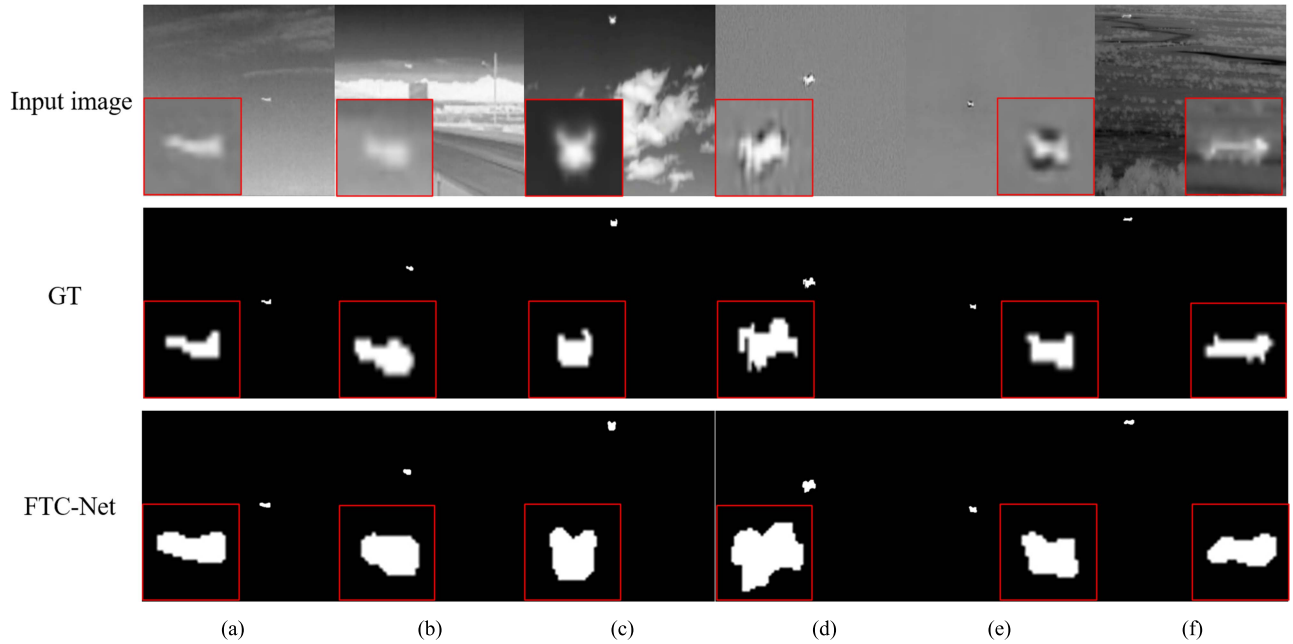

Fig. 9. Samples of inaccurate detection results using the proposed FTC-Net.

In addition, there are often a large number of false alarm areas, as shown in Fig. 6(a) and 6(b). It can be explained from two aspects. First, traditional models tend to focus on the differences between the small target and the background environment, which is not suitable for detecting targets in dark environments. Second, some strong clutter signals may be as sparse as target signals, resulting in a higher false alarm rate. From Figs. 6 and 7, we can conclude that the CNN-based methods perform better than the conventional methods in terms of detection accuracy and are less prone to higher false alarm cases. Moreover, compared with ALCNet, our FTC-Net did not miss the target in Fig. 6(c) and did not generate false detections in Fig. 6(a). For detection accuracy, our FTC-Net has a more accurate segmentation of the target contour. These qualitative results illustrated that the designed dual-path network is well adapted to the challenges of various complex backgrounds, target shapes, and sizes, thus showing better segmentation performance.

E. Ablation Study

To explore whether each component is helpful to model performance, we conduct ablation studies on the SIRST dataset. Specifically, we analyze the detection results by removing every single module.

As shown in Table II, "CNN," "Transformer," and "FTC-Net" denote the CNN branch, transformer branch, and proposed method with feature fusion module, respectively. Compared with the CNN branch, the detection performance is better with the addition of the transformer branch, and the IoU value increases from 74.09% to 76.69%. After adding FFM, the value of IoU is increased by 1.03%. Compared to the baseline, our proposed FTC-Net improves the values of IoU, nIoU, and Pd by 3.63%, 1.36%, and 2.72%, respectively. The results of the ablation study illustrate the effectiveness of feature fusion of two branches. As for runtime, the proposed FTC-Net takes about 0.313s to test on a 384 $\times$ 384 image, which is slightly higher than a single Transformer branch or CNN branch.

The excellent detection capability can be attributed to two perspectives. On the one hand, the transformer with a self-attentive mechanism helps to capture long-range dependencies and thus achieves more accurate segmentation. On the other hand, our results reveal that the FFM can fuse local details and global semantic features, eliminating ambiguity generated with decoder features.

F. Visualization of Feature Maps

To better understand and illustrate the effectiveness of FTC-Net, a visualization of feature maps is presented. Grad-CAM [38] intuitively displays the feature maps learned by the network in the form of a heatmap. Grad-CAM can help us analyze the focus area of the network for a certain category. It performs reverse propagation by selecting the node with the largest softmax value and using the average value of the gradient as the weight. The weights of all the corresponding categories of the feature maps are obtained and then the weighted sum is made.

As shown in Fig. 8, we apply the Grad-CAM method to visualize feature maps of our FTC-Net and CNN-based network. To make a better comparison, we choose the deep feature graph of the network to output heatmaps. It can be seen that the CNN-based network is easy to be disturbed by the background with a similar size and brightness to the target. As shown in Fig. 8(a) and (d), the CNN-based network generates false attention in the background with strong noise and multiclutter. However, the feature map from the FTC-Net is more sensitive to the target, which generates accurate shape segmentation.

G. Error Diagnosis

In this section, we analyze some inaccurate detection results on the SIRST testing dataset. The results in Table I show that the detection probability of our proposed FTC-Net is already quite high and achieves good performance. However, there are also false positives and false negatives, as shown in Fig. 9, the segmentation errors mostly come from some incorrectly predicted pixels which distribute around the target boundary.

False detections occur for two reasons. First, the infrared images in the SIRST dataset are used to capture airborne moving targets far away. As a result, the target edges are blurred, and the difference between the targets and background is small in the image. The segmentation results are prone to generate errors at the image boundary. The second reason that affects detection results is that the ground truth is manually labeled. There are visual biases and ambiguous pixels for the actual images, which influence the training process of the proposed model. Besides, for the small target with around 3×3 size, each pixel error will have a great impact on the final detection results.

V. Conclusion

Precise shape segmentation is the key point of infrared small target detection. In this work, we propose an infrared small target detection network named FTC-Net. Different from existing target detection methods based on CNN, the proposed FTC-Net contains a hierarchical transformer branch to capture long-range contextual dependencies between the targets and background. In addition, to address the feature inconsistency between the transformer-based and CNN-based branch outputs, we design a FFM that can well concatenate long-range contextual information and local edge details. We conduct ablation experiments to illustrate the effectiveness of the transformer branch and feature fusion module. Moreover, qualitative and quantitative results on the SIRST dataset show that the proposed approach achieves high-quality predictions with favorable detection performance and strong generalization ability.

References

[1] M. Teutsch and W. Krüger, "Classification of small boats in infrared images for maritime surveillance," in *Proc. Int. WaterSide Secur. Conf.*, 2010, pp. 1–7.

[2] W. Zhang, M. Cong, and L. Wang, "Algorithms for optical weak small targets detection and tracking: Review," in *Proc. IEEE Int. Conf. Neural Netw. Signal Process.*, vol. 1, 2003, pp. 643–647.

[3] F. S. Marvasti, M. R. Mosavi, and M. Nasiri, "Flying small target detection in IR images based on adaptive toggle operator," *IET Comput. Vis.*, vol. 12, no. 4, pp. 527–534, 2018.

[4] T. S. Anju and N. R. N Raj, "Shearlet transform based image denoising using histogram thresholdingt," in *Proc. IEEE Int. Conf. Commun. Syst. Netw.*, 2016, pp. 162–166.

[5] X. Y. Wang, Z. M. Peng, P. Zhang, and Y. M. He, "Infrared small target detection via nonnegativity-constrained variational mode decomposition," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1700–1704, Oct. 2017.

[6] C. L. P. Chen, H. Li, Y. T. 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.

[7] H. Deng, X. P. 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.

[8] X. Bai and Y. Bi, "Derivative entropy-based contrast measure for infrared small-target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2452–2466, Apr. 2018.

[9] J. Han, S. Liu, G. Qin, Q. Zhao, H. Zhang, and N. Li, "A local contrast method combined with adaptive background estimation for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1442–1446, Sep. 2019.

[10] Y. J. He, M. Li, J. L. Zhang, and Q An, "Small infrared target detection based on low-rank and sparse representation," *Infrared Phys. Technol.,* vol. 68, pp. 98–109, 2015.

[11] 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.

[12] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal 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.

[13] X. Wang, Z. Peng, D. Kong, and Y. He, "Infrared dim and small target detection based on stable multisubspace learning in heterogeneous scene," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5481–5493, Oct. 2017.

[14] M. Liu, H.-Y. Du, Y.-J. Zhao, L.-Q. Dong, and M. Hui, "Image small target detection based on deep learning with SNR controlled sample generation," *Curr. Trends Comput. Sci. Mech. Autom.*, vol. 1, pp. 211–220, 2018.

[15] Z. Fan, D. Bi, X. Lei, M. Shiping, L. He, and W. Ding, "Dim infrared image enhancement based on convolutional neural network," *Neurocomputing*, vol. 272, pp. 396–404, 2018.

[16] 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.

[17] 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.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[19] H. Zhu, J. Zhang, G. Xu, and L. Deng, "Balanced ring top-hat transformation for infrared small-target detection with guided filter kernel," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 5, pp. 3892–3903, Oct. 2020.

[20] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Infrared small-target detection using multiscale gray difference weighted image entropy," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 1, pp. 60–72, Feb. 2016.

[21] J. Han, K. Liang, B. Zhou, X. Zhu, J. Zhao, and L. Zhao, "Infrared small target detection utilizing the multiscale relative local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 612–616, Apr. 2018.

[22] H. Wang, C. Liu, C. Ma, and S. Ma, "A novel and high-speed local contrast method for infrared small-target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1812–1816, Oct. 2020.

[23] Y. He, C. Zhang, T. Mu, T. Yan, Y. Wang, and Z. Chen, "Multiscale local gray dynamic range method for infrared small-target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 10, pp. 1846–1850, Oct. 2021.

[24] 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.

[25] K. Wang, S. Li, S. Niu, and K. Zhang, "Detection of infrared small targets using feature fusion convolutional network," *IEEE Access*, vol. 7, pp. 146081–146092, 2019.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[27] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure aware positional transformer for visible-infrared person Re-identification," *IEEE Trans. Image Process.*," vol. 31, pp. 2352–2364, 2022.

[28] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-effifiifficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[29] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 15908–15919, 2021.

[30] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[31] L. Yuan et al., "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 558–567.

[32] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
[33] R. Guo, D. Niu, L. Qu, and Z. Li, "SOTR: Segmenting objects with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7157–7166.
[34] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[35] 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.
[36] Y. Qin, L. Bruzzone, C. Gao, and B. Li, "Infrared small target detection based on facet kernel and random walker," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7104–7118, Sep. 2019.
[37] 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.
[38] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[39] Y. Dai, Y. Wu, Y. Song, and J. Guo, "Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values," *Infrared Phys. Technol.*, vol. 81, pp. 182–194, 2017.
[40] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sens.*, vol. 11, no. 4, pp. 382, Jan. 2019.
[41] 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.
[42] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[43] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Cham, Switzerland: Springer, 2021, pp. 14–24.
[44] E. Michaelsen and U. Stilla, "Estimating urban activity on high-resolution thermal image sequences aided by large scale vector maps," in *Proc. IEEE/ISPRS Joint Workshop Remote Sens. Data Fusion Urban Areas*, 2001, pp. 25–29.
[45] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

Meibin Qi received the B.E. degree in radio technology from Chongqing University, Chongqing, China, in 1991, the M.E. and Ph.D. degrees in signal and information processing from the Hefei University of Technology, Hefei, China, in 2001 and 2007.

He is currently a Professor with the School of Computer and Information, Hefei University of Technology. His research interests include pattern recognition, video coding, video surveillance, and the application of DSP technology.

Liu Liu is currently studying for a master's degree in electronic information with the School of Computer and Information, Hefei University of Technology, Hefei, China.

Her current research interests include computer vision and object detection.

Shuo Zhuang received the B.E. degree in electrical engineering and automation from Tianjin University, Tianjin, China, in 2020, the M.E. and Ph.D. degrees in pattern recognition and intelligent system from Tianjin University, in 2020.

He is currently a Lecturer with the School of Computer and Information, Hefei University of Technology, Hefei, China. His research interests include pattern recognition, object detection, and object tracking.

Yimin Liu is currently studying for his Ph.D. degree in information and communication engineering with the School of Computer and Information, Hefei University of Technology, Hefei, China.

His current research interests include computer vision and pedestrian reidentification.

Kunyuan Li is currently studying for his Ph.D. degree in information and communication engineering with the School of Computer and Information, Hefei University of Technology, Hefei, China.

His current research interests include computer vision and image processing.

Yanfang Yang received the B.E. degree in applied physics from Fudan University, Shanghai, China, in 1991, the M.E. degree in curriculum and teaching theory from the Hefei University of Technology, Hefei, China, in 2003.

She is currently an Associate Professor with the School of Physics, Hefei University of Technology.

Xiaohong Li is currently an Associate Professor with the School of Computer and Information, Hefei University of Technology, Hefei, China. Her current research interests include computer vision and pedestrian reidentification.