2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

FDFE-Net: Frequency Domain Feature Enhancement Network for Infrared Small Target Detection

Haoyu Zuo, Xincheng Zhang, Zhou Yang, Jiazhen Huang, and Xu Wang(✉)

College of Computer Science, Sichuan University, Chengdu 610065, China
wxu@scu.edu.cn

Abstract. Infrared small target detection (IRSTD) encounters challenges due to the tiny sizes of targets and interference from complex backgrounds. To overcome these issues, this paper proposes a novel Frequency Domain Feature Enhancement Network (FDFE-Net). The proposed network significantly improves the detection accuracy and robustness for IRSTD by integrating the micro-scale feature encoder (MSF Encoder) and frequency domain feature enhancement (FDFE) module. Specifically, the MSF Encoder combines parallel feature extraction and feature enhancement modules to effectively capture multi-scale feature information, thus mitigating information loss. The FDFE module introduces frequency domain features via the Haar wavelet transform, enhancing the semantic differences between targets and backgrounds, thereby improving the distinguishability of small targets. Experimental results on three public datasets, NUAA-SIRST, NUDT-SIRST, and IRSTD-1K, demonstrate that the proposed FDFE-Net outperforms several state-of-the-art IRSTD methods across multiple evaluation metrics.

Keywords: Infrared small target detection, Deep learning, Frequency domain feature, Haar wavelet transform.

1 Introduction

Infrared small target detection (IRSTD) aims to detect small targets from complex infrared backgrounds. Currently, this technology has been widely applied in various fields such as maritime surveillance [1], infrared tracking [2], and military security [3]. Compared with visible image-based target detection, infrared small targets possess the following characteristics. First, they occupy extremely limited pixels, accounting for less than 0.15% of the entire image. For instance, in an image of 128×128 pixels, the sizes of these small targets typically range from 1×1 to 6×6 pixels [4]. Due to their small size, infrared small targets lack distinct texture and shape features. Second, infrared images contain complex and variable backgrounds, including diverse environments such as buildings, oceans, clouds, and land. Under these background conditions, infrared small targets frequently face challenges of low contrast, making them easily submerged within complex backgrounds and difficult to detect accurately. To tackle the above challenges, various methods have been proposed, primarily categorized into model-driven and data-driven approaches. Traditional model-driven methods include

morphological filtering-based methods [5, 6], local contrast-based methods [7, 8], and low-rank sparse decomposition methods [9, 10, 11]. However, these traditional methods often require manual feature selection and design, which limits their performance in low-contrast conditions and complex scenarios, resulting in poor generalization capability. Data-driven methods [12-15] have introduced deep learning into IRSTD, particularly U-shaped neural networks composed of encoders, decoders, and long-range skip connections, which have been widely adopted.

Although deep learning methods significantly improve detection accuracy, their encoders typically use fixed-size convolution kernels, resulting in limited receptive fields and insufficient encoding of global semantic information. Additionally, due to the small sizes of infrared targets and interference from complex backgrounds, these methods may struggle to distinguish small targets from backgrounds. Furthermore, the details and crucial information regarding small targets is easily lost during multiple downsampling operations in U-shaped neural networks.

To address these issues in IRSTD, we propose a novel Frequency Domain Feature Enhancement Network (FDFE-Net). As illustrated in Fig. 1, FDFE-Net integrates two innovative modules: the micro-scale feature encoder (MSF Encoder) and frequency domain feature enhancement (FDFE) module. Specifically, MSF Encoder combines parallel feature extraction [16] and feature enhancement modules, effectively capturing multi-scale feature information to better address information loss. Additionally, the feature enhancement module employs a multi-branch convolutional structure [17], expanding the receptive field via dilated convolution to extract richer contextual information. FDFE incorporates frequency domain features via the Haar wavelet transform [18], which exhibits the most compact spatial support and optimal edge-matching filtering properties [19]. These characteristics render it particularly suitable for IRSTD. Additionally, FDFE effectively integrates high-dimensional and low-dimensional features. This architecture enables the model to better understand the relationship between targets and backgrounds.

In summary, the contributions of this paper can be summarized as follows:

- We propose FDFE-Net, which introduces frequency domain features via the FDFE module to effectively enhance the semantic differences between targets and backgrounds, thereby improving detection performance and robustness.
- We design an MSF Encoder that integrates parallel feature extraction and feature enhancement modules to effectively capture both local and global contextual information, further enhancing the representation capability for small target features.
- We evaluate the proposed FDFE-Net on three publicly available single-frame infrared image datasets, demonstrating its significant advantages over multiple state-of-the-art IRSTD methods.

2 Related Work

First, we briefly review IRSTD techniques based on traditional methods and deep learning methods. Next, we discuss the application of wavelet transform in image processing.

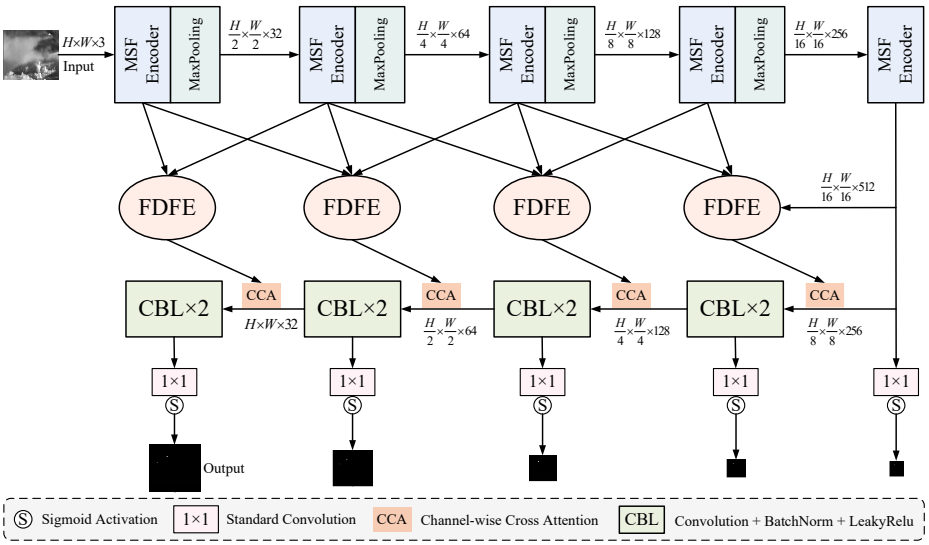

Fig. 1. Overview of the proposed FDFE-Net for infrared small target detection. It is built upon the U-Net architecture and consists of several key modules: MSF Encoder and FDFE module.

2.1 Traditional Methods

IRSTD algorithms can be broadly categorized into model-driven and data-driven methods. Early model-driven algorithms were primarily designed based on prior knowledge to construct filters or modules. For example, Deshpande et al. [5] investigated the maximum mean and maximum median filters to enhance the target regions. TopHat [6] estimates the background using different types of filters. Additionally, inspired by the human visual system, several local contrast-based methods have been proposed. For instance, Wei et al. [7] introduced a multi-scale block-based contrast measurement approach. Aghaziyarati et al. [8] proposed a local contrast measurement method based on the mean absolute gray difference to reduce the missed detection rate. Low-rank sparse decomposition methods have been proposed to handle complex and rapidly changing backgrounds. Examples include Gao et al. [9], who addressed IRSTD using low-rank matrix recovery techniques. Wang et al. [10] integrated variational regularization and principal component pursuit to model background characteristics. However, these models rely on prior knowledge and are sensitive to model parameters and scene variations, resulting in limited generalization capability, particularly in complex environments and low-contrast conditions.

2.2 Deep Learning Methods

In recent years, with the development of deep learning in the field of computer vision, researchers have introduced deep learning into IRSTD to enable automatic feature extraction and efficient detection. These include methods based on Generative Adversarial Network (GAN) and encoder-decoder architectures. GAN-based methods employ

adversarial learning between the generator and discriminator. For example, Wang et al. [12] formulate image segmentation as an optimization problem within the framework of GAN. However, due to the difficulty in achieving an optimal balance during the training process, methods based on GAN are prone to model collapse. On the other hand, encoder-decoder-based methods have gained increasing attention due to their simple structure and training process. For example, asymmetric contextual modulation (ACM) network [13] achieves semantic exchange between high-level and low-level features. Similarly, DNA-Net [14] achieves multi-level feature fusion and adaptive feature enhancement through its dense nested interaction module and spatial attention module. Additionally, Wu et al. [15] modeled IRSTD as a semantic segmentation problem and proposed a straightforward IRSTD framework called U-Net in U-Net (UIU-Net). Compared with traditional methods, the above deep learning approaches have achieved satisfactory performance. However, these algorithms still have some limitations. CNN-based encoders typically use fixed-size convolution kernels, resulting in a limited receptive field that cannot fully capture the global correlation. Moreover, in long-range skip connections, simple skip connections and dense nested modules prove insufficient to enhance the favorable responses of features to the decoder.

2.3 Wavelet Transform in Image Processing

Wavelet transform is a fundamental technique in digital signal processing. It converts images into the frequency domain through a set of spatial filters [20]. This approach has been widely employed to improve the feature representation of image signals, such as in denoising, compression, and super-resolution tasks [21, 22, 23]. In the field of image segmentation, recent studies have investigated wavelet transform applications [24, 25]. In IRSTD tasks, small targets have small sizes, weak thermal signals, and blurry contours, which are often difficult to clearly identify. Therefore, introducing frequency domain features to complement the deficiencies of spatial features and improve the distinguishability of small targets is crucial.

3 Method

3.1 Overall Architecture

As illustrated in Fig. 1, the overall architecture of FDFE-Net follows a U-shaped encoder-decoder architecture. For the given infrared image, FDFE-Net initially utilizes the MSF Encoder and max pooling layers to extract hierarchical features. Then, the output from the encoder is processed by the FDFE module which performs frequency domain feature enhancement and multi-scale feature fusion. Detailed descriptions of the MSF Encoder and FDFE module are provided in Section 3.2 and 3.3, respectively. Subsequently, we employ Channel-wise Cross Attention (CCA) [26] to fuse high-level and low-level features, followed by two CBL blocks for decoding. Finally, the deep supervision strategy [27] is implemented to compute the loss between the overall saliency map and the ground truth label Y, as detailed in Section 3.4.

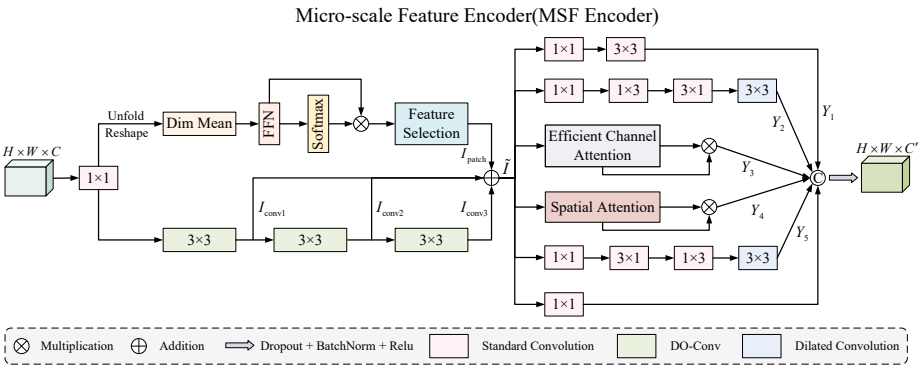

Fig. 2. The structure of the MSF Encoder. This module primarily consists of two components: parallel feature extraction and feature enhancement.

3.2 Micro-scale Feature Encoder

In the IRSTD task, multiple downsampling operations often lead to the loss of fine details and critical target information. Moreover, complex infrared backgrounds can cause false alarms with similar features, exacerbating detection challenges. To address these issues, we propose the MSF Encoder, which combines parallel feature extraction and feature enhancement modules. This architecture enables the MSF Encoder to effectively preserve hierarchical semantic representations across scales, thus better mitigating information loss.

Parallel Feature Extraction. As shown in Fig. 2, the MSF Encoder employs a parallel feature extraction method inspired by HCF-Net [16]. The difference lies in the fact that we introduce the depthwise over-parameterized convolutional layer (DO-Conv) [28] in the sequence convolution process, which further enhances the feature representation capability. This parallel architecture effectively extract multi-scale features of the target, thereby improving the accuracy of small target detection. Specifically, this proposed method performs feature extraction through two parallel branches: the patch-aware and sequence convolution branch. For the input feature map $I \in \mathrm{R}^{H\times W\times C}$, it is initially adjusted through pointwise convolution to obtain $I' \in \mathrm{R}^{H\times W\times C'}$. Subsequently, through these two branches, $I_{\text{patch}} \in \mathrm{R}^{H\times W\times C'}$ and $I_{\text{conv}} \in \mathrm{R}^{H\times W\times C'}$ are computed separately. Finally, the two results are combined to obtain $\tilde{I} \in \mathrm{R}^{H\times W\times C'}$.

First, we apply unfolding and reshaping operations to partition I' into a set of contiguous patches $(p\times p, H/p, W/p, C')$. Next, channel-level averaging is performed on these patches to obtain $(p\times p, H/p, W/p)$, which is then transformed linearly through a feed-forward network (FFN) [29]. Subsequently, we apply an activation function to introduce non-linearity to the spatial dimensions of the feature map, producing a probability distribution of the features, which optimizes the final result by adjusting corresponding weights.

Then, we employ feature selection [30] to select task-relevant features from the tokens and channels. Specifically, let $d = \frac{H \times W}{p \times p}$, and represent the previously weighted output as $(x_i)_{i=1}^{C'}$, where $x_i \in \mathbb{R}^d$ denotes the i-th output token. Feature selection is performed on each token, yielding $\hat{x}_i = Q \cdot sim(x_i, \omega) \cdot x_i$, where $\omega \in \mathbb{R}^{C'}$ and $Q \in \mathbb{R}^{C' \times C'}$ are task-specific parameters, and sim(·,·) denotes the cosine similarity function with a value ranging from 0 to 1. Here, ω serves as the task embedding, indicating the relative importance of each token. Each token x_i is weighted according to its correlation with the task embedding ω (measured by cosine similarity), thereby mimicking the token selection process. Next, a task-specific linear transformation matrix Q is applied for channel selection of each token. Subsequently, reshaping and interpolation operations are performed to obtain the final feature $I_{\text{patch}} \in \mathrm{R}^{H \times W \times C'}$.

In the sequence convolution branch, the input features undergo three consecutive 3×3 DO-Conv [28]. DO-Conv first applies a depthwise convolution kernel $D \in \mathbb{R}^{(H \times W) \times D_{\text{mul}} \times C'}$ to the input feature map, followed by a standard convolution kernel $W \in \mathbb{R}^{C_{\text{out}} \times D_{\text{mul}} \times C'}$, where $D_{mul} = H \times W$ and $C_{out} = C'$. Finally, three separate convolution outputs $I_{\text{conv1}} \in \mathrm{R}^{H \times W \times C'}$, $I_{\text{conv2}} \in \mathrm{R}^{H \times W \times C'}$, and $I_{\text{conv3}} \in \mathrm{R}^{H \times W \times C'}$ are obtained and then summed to produce the final sequence convolution output $I_{\text{conv}} \in \mathrm{R}^{H \times W \times C'}$.

Feature Enhancement. Due to the complexity of infrared images, false alarms with similar features frequently occur in IRSTD tasks. To address this challenge, after feature extraction, we introduce a multi-branch convolutional structure integrated with attention mechanisms to enhance the features of small targets, establishing feature interactions between local and global contexts, thereby reducing the false alarm rate. Each convolutional branch applies a 1 × 1 convolution operation on the input feature map $\tilde{I} \in \mathrm{R}^{H \times W \times C'}$ to preliminarily adjust the channel dimensions required for subsequent processing. The final branch implements a residual structure, which forms an equivalent mapping, thereby preserving critical information about small targets. The remaining three convolutional branches perform cascaded convolution operations, including standard convolution and dilated convolution with a dilation rate of 5, producing outputs Y_1, Y_2, and Y_5. In the third and fourth branches, $\tilde{I}$ is processed through efficient channel attention [31] and spatial attention mechanisms [32], producing a one-dimensional channel attention map $M_c \in \mathrm{R}^{1 \times 1 \times C'}$ and a two-dimensional spatial attention map $M_s \in \mathrm{R}^{H \times W \times 1}$, thereby enhancing the model's feature representation capability in both spatial and channel dimensions. Subsequently, the calculation proceeds according to the following formulas:

$$Y_3 = M_c(\tilde{I}) \otimes \tilde{I},\ Y_4 = M_s(\tilde{I}) \otimes \tilde{I} \tag{1}$$

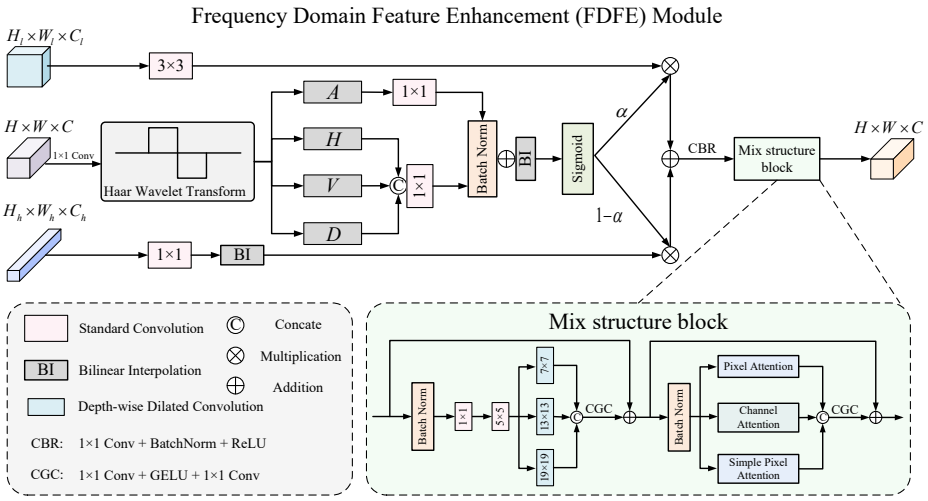

Fig. 3. The structure of the FDFE Module.

$$Y = (\mathrm{Cat}(Y_1, Y_2, Y_5) \oplus f_{conv}^{1\times1}(\tilde{I})) \oplus Y_3 \oplus Y_4 \tag{2}$$

$$Y' = \delta(\mathrm{B}(\mathrm{dropout}(Y))) \tag{3}$$

where $f_{conv}^{k\times k}$ represents the standard convolution operation with a kernel size of $k \times k$. $Cat(\cdot)$ denotes the feature map concatenation operation, $\oplus$ represents element-wise addition of feature maps, and $\otimes$ denotes element-wise multiplication. $Y_c \in \mathrm{R}^{H\times W\times C'}$ and $Y_s \in \mathrm{R}^{H\times W\times C'}$ represent the features after channel and spatial selection, respectively. $\delta(\cdot)$ and $\mathrm{B}(\cdot)$ represent the rectified linear unit (ReLU) and batch normalization functions, respectively. $Y' \in \mathrm{R}^{H\times W\times C'}$ is the final output of the MSF Encoder.

3.3 Frequency Domain Feature Enhancement Module

Due to weak thermal signals, indistinct contours of small targets, and the limited grayscale dynamic range and low contrast of infrared images, small targets are easily overwhelmed by complex backgrounds. As shown in Fig. 3, To address these issues, we propose the FDFE module. FDFE enhances the skip connections within U-Net and employs the Haar wavelet transform [18] to extract frequency domain features. The Haar wavelet transform, owing to its the most compact spatial support and optimal edge-matching filtering properties, efficiently extracts local contour features of small targets while suppressing background noise [19]. Its extracted high-frequency components emphasize edge delineation and detailed features of small targets, whereas the low-frequency components preserve global contextual information of the background. This multi-scale representation effectively compensates for spatial feature limitations, thereby enhancing the model's target-background discrimination capability.

Subsequently, FDFE effectively integrates high-dimensional and low-dimensional features, retaining detailed information of small targets and providing sufficient contextual information. Finally, the Mix Structure Block [33] is employed to further extract and enhance features after multi-scale fusion.

Specifically, FDFE first applies a pointwise convolution to the current layer's feature $i_u \in \mathbb{R}^{H\times W\times C}$, generating a new feature $I_u \in \mathbb{R}^{H\times W\times C}$ while preserving its original dimensions. Subsequently, the Haar wavelet transform is employed to decompose the feature map into four components: a low-frequency component A, a horizontal high-frequency component H, a vertical high-frequency component V, and a diagonal high-frequency component D. The three high-frequency components (H,V,D) are then concatenated and processed through a pointwise convolution for low-dimensional mapping to obtain the high-frequency features. Finally, the high-frequency and low-frequency features are combined through addition and interpolation operations to obtain the frequency domain features $I'_u \in \mathbb{R}^{H\times W\times C}$:

$$I_l, I_h = f_w(i) = (\mathrm{B}(f_{conv}^{1\times1}(A)),\ \mathrm{B}(f_{conv}^{1\times1}(\mathrm{Cat}(H,V,D)))) \tag{4}$$

$$I'_u = B(I_l \oplus I_h) \tag{5}$$

Here, $I_l \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}, I_h \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ represent the low-frequency and high-frequency features, respectively. $\mathrm{B}(\cdot)$ represents batch normalization, and $B(\cdot)$ denotes bilinear interpolation.

Through convolution and interpolation operations, the high-dimensional feature $I_H \in \mathbb{R}^{H_h\times W_h\times C_h}$ and low-dimensional feature $I_L \in \mathbb{R}^{H_l\times W_l\times C_l}$ are aligned with the current layer's feature $I'_u \in \mathbb{R}^{H\times W\times C}$, resulting in $I'_H \in \mathbb{R}^{H\times W\times C}$ and $I'_L \in \mathbb{R}^{H\times W\times C}$. The resulting features are then computed according to the following formula:

$$\alpha = sigmoid(I'_u) \tag{6}$$

$$\tilde{I}_u = \delta(\mathrm{B}(f_{conv}^{1\times1}(\alpha I'_L \oplus (1-\alpha)I'_H))) \tag{7}$$

where $\alpha \in \mathbb{R}^{H\times W\times C}$ is obtained by applying a sigmoid activation function to I'_u. The final enhanced feature map is denoted as $\tilde{I}_u \in \mathbb{R}^{H\times W\times C}$.

Finally, we employ a transformer-style Mix Structure Block [33] for further feature extraction and enhancement of the multi-scale fused features. Specifically, the input feature map $\tilde{I}_u$ is initially processed by batch normalization, yielding the transformed feature map $\hat{I}_u$. The subsequent calculations are conducted as follows:

$$I_{u1} = f_{conv}^{5\times5}(f_{conv}^{1\times1}(\hat{I}_u)) \tag{8}$$

$$I_{u2} = Cat(f_{DWDConv}^{19\times19}(I_{u1}), f_{DWDConv}^{13\times13}(I_{u1}), f_{DWDConv}^{7\times7}(I_{u1})) \tag{9}$$

$$F = \tilde{I}_u \oplus f_{conv}^{1\times1}(GELU(f_{conv}^{1\times1}(I_{u2}))) \tag{10}$$

Here, $f_{DWDConv}^{k\times k}$ denotes a depthwise dilated convolution with dilation rate of 3 which dilated convolution kernel size of $k\times k$. $GELU(\cdot)$ represents the application of the GELU activation function. Subsequently, we process the feature map $F \in \mathbb{R}^{H\times W\times C}$ by integrating multiple attention mechanisms. After applying batch normalization to obtain $\hat{F}$, pixel attention, channel attention, and simple pixel attention are sequentially employed to derive attention-enhanced features F_p, F_c, and F_s. These three attention outputs are then concatenated along the channel dimension. The concatenated features are processed through a Multi-Layer Perceptron (MLP) to align their channel dimension with the input feature map F. Finally, the aligned features are summed with F to produce the final output $F' \in \mathbb{R}^{H\times W\times C}$. The detailed computational procedure is as follows:

$$F_p = \hat{F} \otimes sigmoid(f_{conv}^{1\times1}(GELU(f_{conv}^{1\times1}(\hat{F})))) \tag{11}$$

$$F_c = \hat{F} \otimes sigmoid(f_{conv}^{1\times1}(GELU(f_{conv}^{1\times1}(GAP(\hat{F}))))) \tag{12}$$

$$F_s = f_{conv}^{3\times3}(f_{conv}^{1\times1}(\hat{F})) \otimes sigmoid(f_{conv}^{1\times1}(\hat{F})) \tag{13}$$

$$F' = F \oplus f_{conv}^{1\times1}(GELU(f_{conv}^{1\times1}(Cat(F_s, F_p, F_c)))) \tag{14}$$

Here, $GAP(\cdot)$ denotes global average pooling.

3.4 Loss Function

We integrate a deep supervision strategy [27] into FDFE-Net to enhance gradient flow and mitigate information loss of small targets caused by multiple downsampling operations. Specifically, for each decoder output I_i, we apply 1×1 convolution followed by the sigmoid function to obtain saliency maps. Subsequently, these low-resolution saliency maps are upsampled to the original image size, yielding prediction masks $X_i (i = 1,2,3,4)$, and then all saliency maps are fused to obtain X_5. Finally, we compute the binary cross-entropy (BCE) loss between the prediction masks and ground truth (GT) annotations *Y*. The total loss L is obtained by calculating a weighted sum of all loss terms at multiple scales. The formula is as follows:

$$l_i = \mathrm{L}_{BCE}(X_i, Y),\ L = \sum_{i=0}^{5} \lambda_i \cdot l_i \tag{15}$$

Here, $l_i (i = 0,1,2,3,4,5)$ denotes the loss at each scale, and the corresponding loss weight for each scale is set as $\lambda_i = 1\ (i = 0,1,2,3,4,5)$.

4 Experiments

4.1 Experimental Setup

Dataset. In this study, we conducted experiments on three publicly available single-frame infrared image datasets: NUAA-SIRST [13], NUDT-SIRST [14], and IRSTD-1k [34], containing 427, 1327, and 1000 labeled images respectively. To ensure experimental standardization, we utilize the dataset partitioning approach proposed in [27] to divide the datasets into training and test sets.

Evaluation Metrics. To comprehensively evaluate the detection performance of different algorithms, we adopt multiple metrics, including Intersection over Union (IoU), Normalized IoU (nIoU), Mean IoU (mIoU), F-measure, Probability of Detection (Pd), and False Alarm Rate (Fa). Specifically, IoU is a pixel-level evaluation metric defined as the ratio of intersection and union areas between predicted and ground-truth values. The F-measure evaluates missed detections and false alarms at the pixel-level. Pd is defined as the ratio of correctly predicted target pixels to all target pixels. In contrast, Fa denotes the proportion of false alarm pixels to all image pixels.

Implementation Details. We conducted experiments for FDFE-Net on an NVIDIA GeForce RTX 3070 Ti GPU. Our model does not rely on any pre-trained weights for training. Each image is normalized and randomly cropped into 256×256 patches. To avoid overfitting, we augment the training data through random flipping and rotation. The network is optimized using the Adam optimizer [35] with an initial learning rate of 0.001, which is gradually reduced to 1×10^{-5} via the cosine annealing strategy. The batch size and epoch size are set as 32 and 200, respectively.

4.2 Quantitative Results

We compare the proposed FDFE-Net with several state-of-the-art (SOTA) methods, including ACM [13], ALCNet [36], DNA-Net [14], UIU-Net [15] and SCTransNet [27]. To ensure a fair comparison, we retrained these methods using the same training datasets as FDFE-Net and followed their original thresholds. Table 1, Table 2, and Table 3 present the quantitative results of various metrics. FDFE-Net consistently outperforms other methods across all three public datasets in four key metrics: mIoU, nIoU, F-measure, and Pd. This demonstrates that FDFE-Net not only effectively preserves small target details, but also exhibits superior capability in distinguishing targets from background clutter. We also note that FDFE-Net does not achieve the optimal Fa. For example, DNA-Net has a 1.34% lower Fa than ours on the NUAA-SIRST dataset. However, our method surpasses DNA-Net by 2.07% in detection accuracy. This demonstrates that FDFE-Net achieves a favorable trade-off between false alarm rate and detection accuracy. Furthermore, we comprehensively compare FDFE-Net with the most competitive deep learning methods, UIU-Net and SCTransNet. Table 4 presents the average metrics of these methods across the three datasets. We observe that FDFE-Net

consistently outperforms the other methods. This demonstrates that FDFE-Net achieves a dual optimization of computational efficiency and model performance.

Table 1. Comparisons with SOTA methods on NUAA-SIRST.

Methods	mIoU	nIoU	F-measure	Pd	Fa
ACM [13]	68.93	69.18	80.87	91.63	15.23
ALCNet [36]	70.83	71.05	82.92	94.30	36.15
DNA-Net [14]	75.80	79.20	86.24	95.82	**8.78**
UIU-Net [15]	76.91	79.99	86.95	95.82	14.13
SCTransNet [27]	77.50	81.08	87.32	96.95	13.92
FDFE-Net	**78.31**	**83.55**	**88.72**	**97.89**	10.12

Table 2. Comparisons with SOTA methods on NUDT-SIRST.

Methods	mIoU	nIoU	F-measure	Pd	Fa
ACM [13]	61.12	64.40	75.87	93.12	55.22
ALCNet [36]	64.74	67.20	78.59	94.18	34.61
DNA-Net [14]	88.19	88.58	93.73	98.83	9.00
UIU-Net [15]	93.48	93.89	96.63	98.31	7.79
SCTransNet [27]	94.09	94.38	96.95	98.62	**4.29**
FDFE-Net	**94.22**	**94.71**	**97.24**	**99.32**	5.47

Table 3. Comparisons with SOTA methods on IRSTD-1K.

Methods	mIoU	nIoU	F-measure	Pd	Fa
ACM [13]	59.23	57.03	74.38	93.27	65.28
ALCNet [36]	60.60	57.14	75.47	92.98	58.80
DNA-Net [14]	65.90	66.38	79.44	90.91	12.24
UIU-Net [15]	66.15	66.66	79.63	93.98	22.07
SCTransNet [27]	68.03	68.15	80.96	93.27	10.74
FDFE-Net	**68.32**	**68.25**	**81.21**	**94.34**	**10.26**

Table 4. Comprehensive evaluation metrics with competitive algorithms

Methods	Params(M)	FLOPs(G)	mIoU	nIoU	F-measure
UIU-Net [15]	50.540	54.42	78.85	80.18	87.74
SCTransNet [27]	11.190	20.24	79.87	81.20	88.41
FDFE-Net	**10.203**	**18.26**	**80.28**	**82.17**	**89.06**

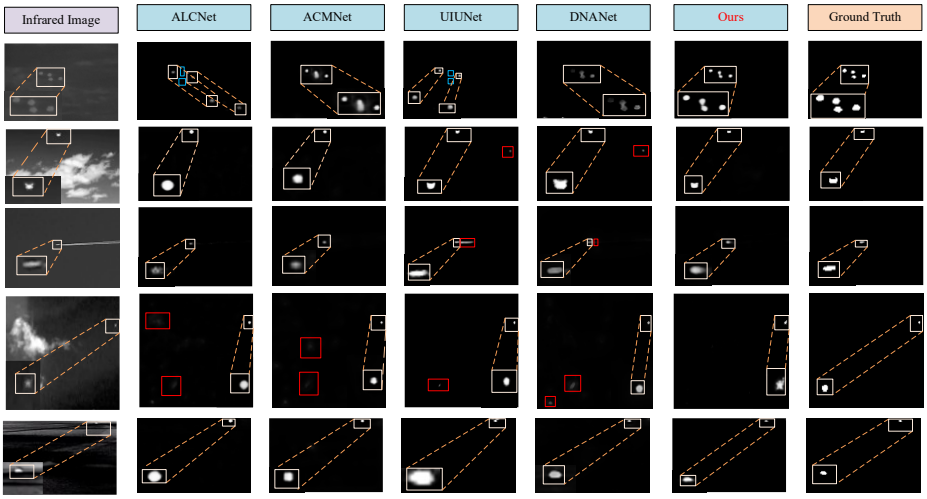

Fig. 4. Visual results of representative methods. White, red, and blue boxes represent correctly detected targets, false alarms, and missed detections respectively.

4.3 Visual Results

Fig. 4 illustrates the visualization results of various methods. Compared to other methods, FDFE-Net achieves superior accurate target detection and contour segmentation. Specifically, as shown in the first row, our method accurately detects more small targets, whereas other methods exhibit missed detections. Rows two to four indicate that in cluttered backgrounds, our method accurately localizes small targets while effectively suppressing false alarms, which is poorly handled by other methods. This is because our approach not only learns the features of small targets but also effectively captures global contextual information within the image. In contrast, other methods are typically confined to extracting local features and fail to effectively model long-range dependencies, which limits their performance in complex scenarios. Additionally, the first and final row demonstrates that FDFE-Net provides detailed description of shape and texture features.

4.4 Ablation Study

To verify the effectiveness of the proposed MSF Encoder and FDFE module, we conducted ablation studies on the NUAA-SIRST dataset. Specifically, we incrementally replace the encoder and long-range skip connections in the baseline model SCTransNet [27] with our MSF Encoder and FDFE module respectively, then perform comparative evaluations under identical experimental configurations to assess their performance in IRSTD. As shown in Table 5, the experimental results show progressive performance improvements with the incremental incorporation of these modules, confirming the effectiveness of both the MSF Encoder and FDFE module in enhancing detection accuracy and robustness.

Table 5. Ablation study of the MSF Encoder and FDFE module on NUAA-SIRST.

Baseline	MSF Encoder	FDFE	mIoU	nIoU	F-measure	Pd	Fa
√	-	-	77.50	81.08	87.32	96.95	13.92
√	√	-	77.67	81.49	87.42	96.99	14.11
√	-	√	78.11	82.23	87.94	97.22	12.17
√	√	√	**78.31**	**83.55**	**88.72**	**97.89**	**10.12**

5 Conclusion

This paper proposes a novel Frequency Domain Feature Enhancement Network (FDFE-Net) designed to improve the accuracy and robustness of infrared small target detection. By integrating the proposed micro-scale feature encoder (MSF Encoder) and frequency domain feature enhancement (FDFE) module, FDFE-Net significantly enhances small target detection performance, particularly under cluttered backgrounds and low-contrast conditions. Experimental results demonstrate that FDFE-Net outperforms several state-of-the-art methods across three public datasets. Overall, FDFE-Net provides an effective solution for IRSTD and demonstrates its powerful detection capability in complex environments.

References

1. Teutsch, M., Krüger, W.: Classification of small boats in infrared images for maritime surveillance. In: 2010 International WaterSide Security Conference. pp. 1–7 (2010)
2. Huang, Y., Li, X., Lu, R., Hu, Y., Yang, X.: Infrared maritime target tracking via correlation filter with adaptive context-awareness and spatial regularization. Infrared Physics & Technology. **118**, 103907 (2021)
3. Ying, X., Wang, Y., Wang, L., Sheng, W., Liu, L., Lin, Z., Zhou, S.: Local Motion and Contrast Priors Driven Deep Network for Infrared Small Target Superresolution. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. **15**, 5480–5495 (2022)
4. Mao, X., Diao, W.: Criterion to Evaluate the Quality of Infrared Small Target Images. J Infrared Milli Terahz Waves. **30**, 56–64 (2009)
5. Deshpande, S.D., Er, M.H., Venkateswarlu, R., Chan, P.: Max-mean and max-median filters for detection of small targets. In: Signal and Data Processing of Small Targets 1999. pp. 74–83. SPIE (1999)
6. Zeng, M., Li, J., Peng, Z.: The design of top-hat morphological filter and application to infrared target detection. Infrared physics & technology. **48**, 67–76 (2006)
7. Wei, Y., You, X., Li, H.: Multiscale patch-based contrast measure for small infrared target detection. Pattern Recognition. **58**, 216–226 (2016)
8. Aghaziyarati, S., Moradi, S., Talebi, H.: Small infrared target detection using absolute average difference weighted by cumulative directional derivatives. Infrared Physics & Technology. **101**, 78–87 (2019)

9. Gao, C., Wang, L., Xiao, Y., Zhao, Q., Meng, D.: Infrared small-dim target detection based on Markov random field guided noise modeling. Pattern Recognition. **76**, 463–475 (2018)
10. Wang, X., Peng, Z., Kong, D., Zhang, P., He, Y.: Infrared dim target detection based on total variation regularization and principal component pursuit. Image and Vision Computing. **63**, 1–9 (2017)
11. Zhang, L., Peng, L., Zhang, T., Cao, S., Peng, Z.: Infrared small target detection via non-convex rank approximation minimization joint l 2, 1 norm. Remote Sensing. **10**, 1821 (2018)
12. Wang, H., Zhou, L., Wang, L.: Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8509–8518 (2019)
13. Dai, Y., Wu, Y., Zhou, F., Barnard, K.: Asymmetric contextual modulation for infrared small target detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 950–959 (2021)
14. Li, B., Xiao, C., Wang, L., Wang, Y., Lin, Z., Li, M., An, W., Guo, Y.: Dense nested attention network for infrared small target detection. IEEE Transactions on Image Processing. **32**, 1745–1758 (2022)
15. Wu, X., Hong, D., Chanussot, J.: UIU-Net: U-Net in U-Net for infrared small object detection. IEEE Transactions on Image Processing. **32**, 364–376 (2022)
16. Xu, S., Zheng, S., Xu, W., Xu, R., Wang, C., Zhang, J., Teng, X., Li, A., Guo, L.: Hcf-net: Hierarchical context fusion network for infrared small object detection. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
17. Zhang, Y., Ye, M., Zhu, G., Liu, Y., Guo, P., Yan, J.: FFCA-YOLO for small object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. **62**, 1–15 (2024)
18. Yang, Y., Yuan, G., Li, J.: Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. IEEE Transactions on Geoscience and Remote Sensing. (2024)
19. Mashford, J., Rahilly, M., Lane, B., Marney, D., Burn, S.: Edge detection in pipe images using classification of haar wavelet transforms. Applied artificial intelligence. **28**, 675–689 (2014)
20. Fujieda, S., Takayama, K., Hachisuka, T.: Wavelet convolutional neural networks. arXiv preprint arXiv:1805.08620 (2018)
21. Tian, C., Zheng, M., Zuo, W., Zhang, B., Zhang, Y., Zhang, D.: Multi-stage image denoising with the wavelet transform. Pattern Recognition. **134**, 109050 (2023)
22. Akyazi, P., Ebrahimi, T.: Learning-based image compression using convolutional autoencoder and wavelet decomposition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019)
23. Li, W., Guo, H., Liu, X., Liang, K., Hu, J., Ma, Z., Guo, J.: Efficient Face Super-Resolution via Wavelet-based Feature Enhancement Network. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4515–4523. Association for Computing Machinery, New York, NY, USA (2024)
24. Gao, J., Wang, B., Wang, Z., Wang, Y., Kong, F.: A wavelet transform-based image segmentation method. Optik. **208**, 164123 (2020)
25. Arivazhagan, S., Ganesan, L., Bama, S.: Fault segmentation in fabric images using Gabor wavelet transform. Machine Vision and Applications. **16**, 356–363 (2006)
26. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence. pp. 2441–2449 (2022)

27. Yuan, S., Qin, H., Yan, X., Akhtar, N., Mian, A.: Sctransnet: Spatial-channel cross transformer network for infrared small target detection. IEEE Transactions on Geoscience and Remote Sensing. (2024)
28. Cao, J., Li, Y., Sun, M., Chen, Y., Lischinski, D., Cohen-Or, D., Chen, B., Tu, C.: Do-conv: Depthwise over-parameterized convolutional layer. IEEE Transactions on Image Processing. **31**, 3726–3736 (2022)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems. **30**, (2017)
30. Shi, B., Gai, S., Darrell, T., Wang, X.: Refocusing Is Key to Transfer Learning. CoRR. **abs/2305.15542**, (2023)
31. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020)
32. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
33. Lu, L., Xiong, Q., Xu, B., Chu, D.: MixDehazeNet: Mix structure block for image dehazing network. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–10. IEEE (2024)
34. Zhang, M., Zhang, R., Yang, Y., Bai, H., Zhang, J., Guo, J.: ISNet: Shape matters for infrared small target detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 877–886 (2022)
35. Loshchilov, I., Hutter, F., others: Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 (2017)
36. Dai, Y., Wu, Y., Zhou, F., Barnard, K.: Attentional local contrast networks for infrared small target detection. IEEE transactions on geoscience and remote sensing. **59**, 9813–9824 (2021)