HDNet: A Hybrid Domain Network With Multiscale High-Frequency Information Enhancement for Infrared Small-Target Detection

Mingzhu Xu, *Member, IEEE*, Chenglong Yu, Zexuan Li, Haoyu Tang, *Member, IEEE*, Yupeng Hu, *Member, IEEE*, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*— The infrared small-target detection (IRSTD) task involves identifying and separating small targets from complex backgrounds. However, these targets pose significant challenges due to their small, variable sizes and dim appearance with a low signal-to-noise ratio, often obscured by cluttered backgrounds. Standard spatial-domain convolutional neural networks (CNNs) act as low-pass filters, hindering their ability to detect small, variably sized, low-contrast targets against complex backgrounds. Infrared images (IRIs) also exhibit diverse spectral energy distributions, yet CNNs lack a global spectral view to discern these patterns, making them susceptible to background clutter. To address these shortcomings, we propose a novel hybrid-domain network (HDNet), which fuses frequency-domain features with conventional spatial-domain CNN features to markedly enhance target–background contrast and explicitly suppress background interference. Specifically, the HDNet comprises two main branches: the spatial-domain branch and the frequency-domain branch. In the spatial domain, we innovatively introduce a multiscale atrous contrast (MAC) convolution module, utilizing multiple parallel atrous contrast convolutions (ACCs) with varying kernel sizes to enhance the perception of small, variably sized targets. In the frequency domain, we have specifically designed the dynamic high-pass filter (DHPF) module, hierarchically calculating low-frequency signal energy and dynamically removing specific low-frequency information to preserve high-frequency image details. This effectively filters out slowly varying low-frequency backgrounds, highlighting small targets. Comprehensive ablation studies and experimental analysis on three datasets (IRSTD-1K, NUAA-SIRST, and NUDT-SIRST) validate the HDNet's effectiveness and superiority compared to 26 state-of-the-art (SOTA) methods. The source code is available at https://github.com/xumingzhu989/HDNet-TGRS**

***Index Terms*— Dynamic high-pass filter (DHPF), high-frequency information enhancement, infrared small-target detection (IRSTD), multiscale atrous contrast (MAC) convolution.**

Received 4 January 2025; revised 9 March 2025 and 26 April 2025; accepted 22 May 2025. Date of publication 29 May 2025; date of current version 16 June 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62206157 and in part by the Natural Science Foundation of Shandong Province under Grant ZR2022QF047. *(Corresponding author: Yupeng Hu.)*

Mingzhu Xu, Chenglong Yu, Zexuan Li, Haoyu Tang, and Yupeng Hu are with the School of Software, Shandong University, Jinan 250101, China (e-mail: xumingzhu@sdu.edu.cn; yucl@mail.sdu.edu.cn; lzxferb@mail.sdu.edu.cn; tanghao258@sdu.edu.cn; huyupeng@sdu.edu.cn).

Liqiang Nie is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: nieliqiang@gmail.com).

Digital Object Identifier 10.1109/TGRS.2025.3574962

I. Introduction

INFRARED small-target detection (IRSTD) aims to identify the small target and separate it from a complex background in infrared images (IRIs). It boasts a wide range of applications, encompassing military reconnaissance and traffic monitoring [1], [2], [3]. Compared to the generic objects in natural scene images, the small targets in IRIs pose two major challenges: their small, variable sizes, and their dim appearance with a low signal-to-noise ratio. As shown in Fig. 1, due to varying distances between the camera and targets, infrared small targets occupy just a few to dozens of pixels, making accurate separation difficult. Moreover, a large amount of low-frequency background, characterized by their gradual variations (such as the cloud shown in the second row of Fig. 1, exhibiting low brightness and blurred contours), diminishes the visibility and contrast of the intended targets. Consequently, these targets become susceptible to being obscured by complex backgrounds. All these obstacles may result in false alarms or missed targets, rendering IRSTD still a challenging task.

IRSTD can be broadly divided into two important branches. The first, which this article focuses on, is single-frame small-target segmentation [4], [5], [6], where targets are detected by predicting the class of each pixel in the spatial domain. The second is multiframe bounding box detection [7], [8], [9], [10], which leverages motion cues across frames to regress bounding boxes around small targets. We focus on the first single-frame small-target segmentation. Early traditional approaches can be categorized into several types, including filter-based methods [11], [12], [13], local contrast-based methods [14], [15], [16], [17], and low-rank representation methods [18], [19], [20], [21], [22]. These methods have addressed the issues of IRSTD to some extent. For example, the filter-based methods can effectively suppress uniform background noise, while the local contrast-based methods enhance the contrast between the target and the background. However, the large amount of slowly changing low-frequency background areas (such as cloud, fog, or smoke) also overlapped with the high-frequency information of the target, causing these filter-based and contrast-based methods to be ineffective at distinguishing the targets from the complex background. Although the low-rank methods are suitable for low signal-to-noise ratio IRIs, they still tend to produce higher false alarms,

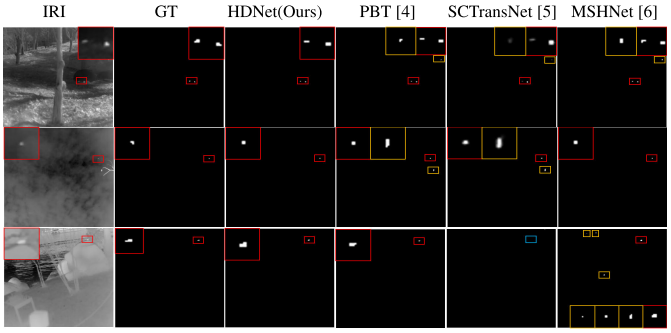

Fig. 1. Visual examples generated by several SOTA methods. The correctly detected targets, false alarms, and missed targets are framed with red, yellow, and blue bounding boxes, respectively. For better visualization, a close-up view of the target is shown in image corners.

leading to unstable detection performance. Furthermore, traditional methods often lack sufficient adaptability and stability when facing varying target scales and environmental changes.

In recent years, with the development of deep learning, especially convolutional neural networks (CNNs), many significant progresses have been made for IRSTD. For example, the ISNet [23] applies Taylor finite differences (TFDs) to capture the shape features of targets and suppress background noise. The DNANet [24] customizes a dense nested interaction module for multilayer feature fusion. The ACMNet [25] proposes an asymmetric context modulation feature fusion method. The RISTDNet [26] proposes a robust infrared small-target detection network. However, these pure CNN-based frameworks, limited by local receptive fields, struggle to capture global context and suppress global noise interference. Their downsampling operations, acting as low-pass filters, often result in missed detection of small, sparsely featured objects, thereby undermining the model's robustness. Furthermore, hybrid CNN–Transformer frameworks have been proposed to overcome the limited global context modeling of pure CNNs. Models like the SCTransNet [5] and PBT [4] leverage computationally intensive self/cross-attention across multilevel CNN features to enhance global representation, improving small-object detection and background suppression. However, IRIs also exhibit diverse spectral energy distributions, while both CNNs and Transformers operate mainly in the spatial domain, lacking a frequency-aware view to distinguish different patterns. As shown in Fig. 1, even advanced models such as the SCTransNet [5] and PBT [4] still struggle with complex background interference. Consequently, recent frameworks such as the FDDBA-NET [27], FDA-IRSTD [28], and HLSR-Net [29] integrate CNNs with frequency-domain processing, extracting features through plain convolutions and suppressing background noise via high/low-frequency decomposition. However, these frameworks typically rely on plain convolutions or predefined thresholds for frequency separation, limiting their adaptability to complex, dynamic low-frequency backgrounds in IRIs.

To address these issues, we propose a novel hybrid-domain network (HDNet), which leverages the perception capability of MAC convolutions in the spatial domain for detecting small- and variable-sized targets, along with the suppression effect of a dynamic high-pass filter (DHPF) in the frequency domain on slowly varying low-frequency backgrounds. Specifically, the HDNet comprises two primary subnetworks: the spatial-domain subnetwork and the frequency-domain subnetwork. In the spatial-domain subnetwork, a classic encoder–decoder architecture is employed. Unlike existing spatial-domain branches that use plain multiscale convolutions, we utilize specially designed multiscale atrous contrast (MAC) convolutions as their basic building blocks, enhancing the perception capability for targets of small and variable sizes. The decoder adopts plain convolutional blocks to progressively upsample and recover the target information and generate the multiscale prediction maps. In the frequency-domain subnetwork, instead of fixed-threshold filtering used in prior works, we propose a novel DHPF to adaptively remove low-frequency components. By iteratively filtering low-frequency information at multiple stages, our approach effectively suppresses slowly varying background interference. Finally, the results from the spatial domain are fused with the results from the frequency domain to obtain the final infrared small-target prediction map.

In summary, our main contributions are as follows.

1) We propose a novel HDNet for the IRSTD task. It takes advantage of the multiscale target perception capability in the spatial domain and the low-frequency information suppression ability in the frequency domain to enhance the performance of IRSTD.
2) We propose a novel MAC convolution module in the spatial domain. This module improves the contrast between targets and cluttered backgrounds, enhancing the perception capability of small and variable-sized targets.
3) We propose a novel DHPF module in the frequency domain. This module calculates the energy of low frequencies and dynamically removes a specific proportion of them, effectively suppressing the slowly varying low-frequency background interference.
4) We have conducted comprehensive ablation studies and experimental analysis on three public datasets (including IRSTD-1K, NUAA-SIRST, and NUDT-SIRST), validating the effectiveness and superiority of our HDNet, compared with 26 state-of-the-art (SOTA) methods.

II. Related Works

In this section, we will discuss the related works on infrared small-target detection and representation learning in the frequency domain.

A. Infrared Small-Target Detection

IRSTD methods have evolved significantly over the years, with approaches broadly categorized into traditional methods and deep learning methods. Traditional methods, such as filter-based approaches [11], [12], [13], leverage spatial- or frequency-domain filters to suppress background noise. While effective for uniform backgrounds, they struggle with complex scenes where target and background frequency components overlap. Low-rank representation-based methods [18], [19], [20], [21], [22], by decomposing images into low-rank (background) and sparse (target) matrices, excel in low-SNR scenarios. However, they often produce high false alarms due

to insufficient discrimination between sparse noise and true targets. Local contrast-based methods [14], [15], [16], [17] enhance target saliency by computing local intensity differences. Although these methods are effective for small targets, their reliance on fixed window sizes limits their adaptability to varying target scales. These traditional methods depend heavily on handcrafted priors and static parameters, making them ineffective for real-world scenes with cluttered backgrounds and multiscale targets. In contrast, recent advances in deep learning have shifted IRSTD toward data-driven paradigms. MDvsFA [30] introduces adversarial learning to balance false alarms and missed detections, but struggles with fine-grained target localization. The ACMNet [25] combines high-level semantics with fine details via asymmetric context modulation, yet lacks explicit mechanisms to suppress low-frequency backgrounds. The ALCNet [31] enhances features by extracting target information using a single local contrast. However, a single local contrast is difficult to adapt to scenes with complex backgrounds and multiscale targets. The UIU-Net [32] integrates nested U-Nets to enhance global–local features, but incurs high computational costs. Dim2Clear [33] learns discriminative features of small infrared targets through mutual guidance between low-level and high-level feature maps. The FTC-Net [34] proposes to leverage the spatial locality of CNNs to preserve small infrared targets as much as possible, while employing Transformers to model long-range dependencies, thereby effectively reducing interference from complex backgrounds and achieving better performance than single-branch models. IRSAM [35] redesigns the general vision segmentation model SAM to address the IRSTD task and achieves advanced performance in most sample scenarios. For multiframe infrared small-target detection, modeling the spatiotemporal context is crucial. The SSTNet [7] introduces a novel cross-slice ConvLSTM nodes to capture motion features within and across slices and designs a motion-coupling neck to exploit multiframe motion cues, achieving superior performance in detecting dim and small moving targets. Tridos [8] employs a triple-domain strategy, enhancing spatiotemporal context with frequency-aware local and global features, and uses residual compensation to mitigate potential cross-domain feature mismatches. To reduce reliance on costly labeled data, S2MVP [9] explores the potential of unlabeled data through a semi-supervised multiview prototype learning framework with motion reconstruction and enhances label quality by eliminating noisy pseudo-labels using an anomaly-driven filtering mechanism. To address the coarse motion cues from pure vision modality, MoPKL [10] leverages homogeneous language descriptions tailored for moving targets and refines motion features via motion–vision alignment and graph attention, boosting moving infrared small-target detection.

These methods hinge on model design for more effective feature extraction. However, the inherent faintness of small targets frequently misleads the model, rendering it vulnerable to interference from complex backgrounds. Additionally, targets of varying scales can trigger false alarms. To tackle these challenges, we have devised an MAC convolution module that captures a broader receptive field across different resolutions. By integrating multiscale features and suppressing local background, this module can precisely extract multiscale targets from complex backgrounds.

B. Representation Learning in the Frequency Domain

The concept of feature learning in the frequency domain involves transforming an image to the frequency domain via the fast Fourier transform (FFT) algorithm, performing a series of operations, and then converting it back to the spatial domain through the inverse FFT (iFFT) algorithm. In recent years, feature learning based on the frequency domain has seen rapid development. The GFNet [36] introduces a global filter layer to enable global information interaction in the frequency domain, learning long-range spatial dependencies with lower time complexity. AFNO [37] modifies the fundamental architecture of FNO [38] by performing token mixing in the frequency domain. Due to the global nature of the frequency domain, this approach addresses the challenge of discontinuity in image representation learning. The AFFNet [39] designs an adaptive frequency band filter operator that achieves success in many vision tasks through global information fusion, theoretically demonstrating that frequency-domain transformation facilitates global integration, offering an efficient, low-power equivalent to global fusion achieved by convolution. SgMg [40] uses frequency-domain guidance for cross-modal fusion, enhancing visual features and promoting global information interaction. The DFFormer [41] achieves variations in filter parameters in the frequency domain by utilizing learnable weights, thereby furnishing a trainable variant of traditional global filters. These methods provide strong theoretical support for the development of feature learning in the frequency domain.

Frequency-domain analysis has also been studied in the task of infrared small-target detection. FDDBA-NET [27] proposes a frequency-domain decoupling mechanism to separate target and background features. FDA-IRSTD [28] introduces a block-level FFT to learn and partition the high- and low-frequency components of each block, thereby suppressing the background. The HLSR-Net [29] proposes a high- and low-frequency semantic reconstruction strategy to suppress the background and extract sparse features of the target. Existing methods often apply static filters or fixed thresholds, which are less effective in handling the varying characteristics of IRIs. Additionally, the inherently large background areas in IRIs often lead to the loss of target information. Unlike these frequency-domain approaches that extract high- and low-frequency features, our DHPF module dynamically adjusts its filtering masks based on the image content, providing more adaptive and context-aware suppression of background noise. By calculating the energy ratio in the frequency domain, we dynamically filter out low-frequency information, thereby suppressing low-frequency background and enhancing foreground representation in IRIs.

III. Proposed Method

In this section, we elaborate on the proposed HDNet. Section III-A provides an overview of the HDNet architecture. Section III-B introduces the design of our MAC

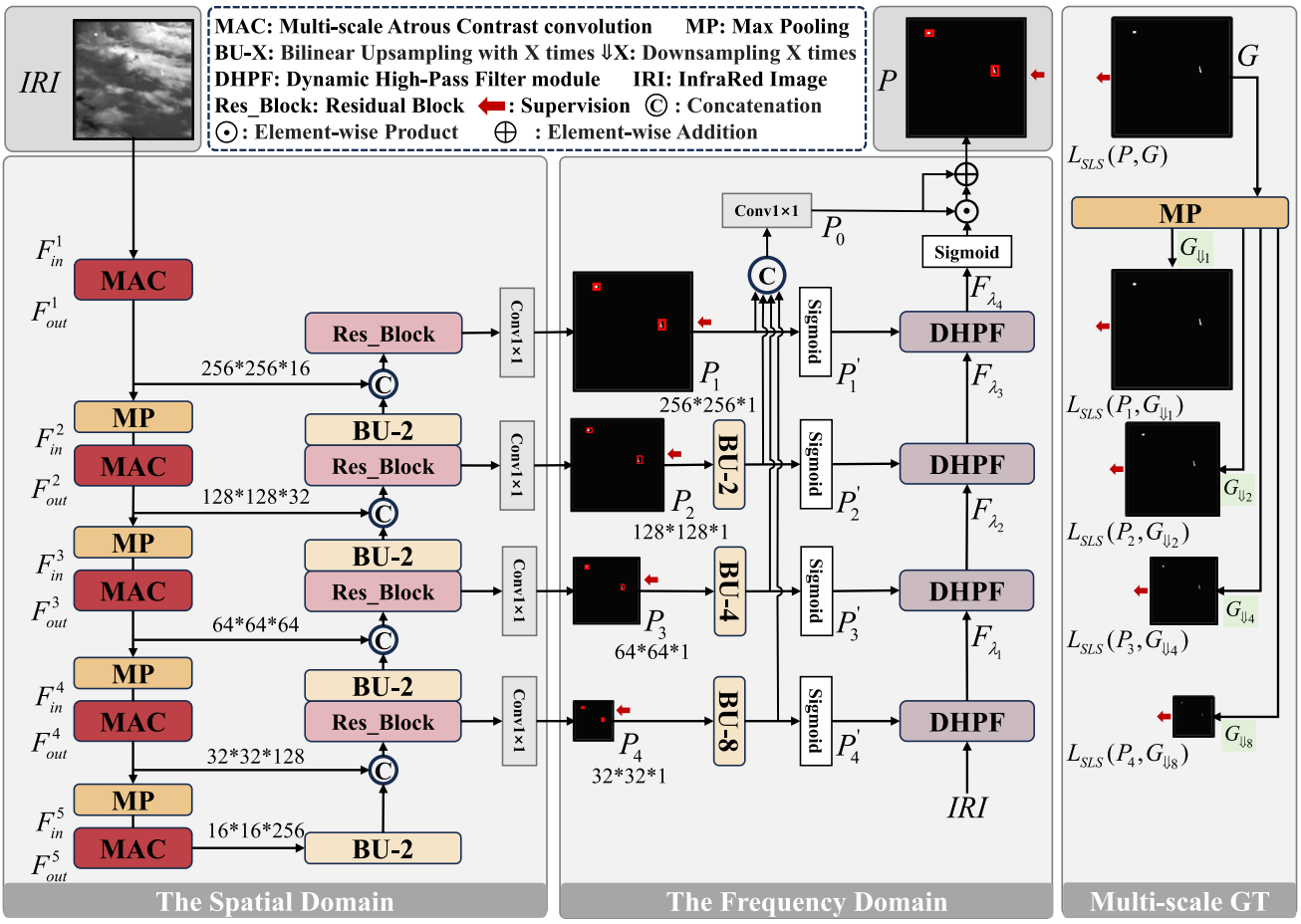

Fig. 2. Overall architecture of our HDNet. It comprises two subnetworks: the spatial-domain subnetwork, which incorporates our novel MAC module at various encoder stages to bolster its ability to perceive multiscale small targets, and the frequency-domain subnetwork, which systematically hierarchically integrates DHPF modules to progressively eliminate slowly varying low-frequency backgrounds, guided by multiscale prediction maps.

module. Section III-C details the DHPF module. Section III-D describes the loss function.

A. Overall Architecture

As illustrated in Fig. 2, our HDNet mainly contains a spatial-domain subnetwork and a frequency-domain subnetwork. **In the spatial domain**, a classic encoder–decoder framework is employed. The encoder uses specially designed MAC convolutions as its core components, while the decoder uses plain convolutional blocks to progressively upsample and recover target information. Specifically, the original IRI is fed into the spatial-domain subnetwork, where multilevel contrast features are extracted through five encoder stages. These features are then fused with the decoder via skip connections, enhancing contextual information capture. Four decoder stages also generate multilevel prediction maps, aiding in the suppression of complex background interference in the following frequency domain. **In the frequency domain**, we integrate specially designed DHPFs at various stages to progressively remove low-frequency information from the image. In the DHPF module, the input (either the original IRI or an intermediate feature map) is first filtered using decoder-stage prediction maps. We then calculate the frequency energy of the internal feature map and progressively reduce low-frequency energy, effectively suppressing slowly varying background information. **In the end**, we merge the final spatial prediction map P_0 with the frequency-domain prediction map to produce the final infrared small-target prediction map P. To further improve the performance of IRSTD, we propose supervising all intermediate prediction maps P_i ($i \in [1, 2, 3, 4]$) and final prediction map P generated by various decoder stages.

B. MAC Block in the Spatial Domain

As shown in Fig. 3(a), infrared small targets typically present small and variable sizes (ranging from one pixel to tens or even hundreds of pixels) and are also with dim appearance and a low signal-to-noise ratio. Although existing CNN methods, such as Res2Net [42], leverage multiconvolution kernel to enhance the perception of targets with variable sizes, they frequently result in information loss when processing small targets with limited textural information, ultimately causing missed targets. To tackle this problem, we propose a novel MAC convolution module in the spatial domain. Unlike Res2Net, our MAC block consists of three parallel ACCs with different kernel sizes. These kernels are specifically designed to amplify the contrast computation between targets of variable sizes and the complex background, enhancing the model's perception capabilities for targets with small and variable sizes.

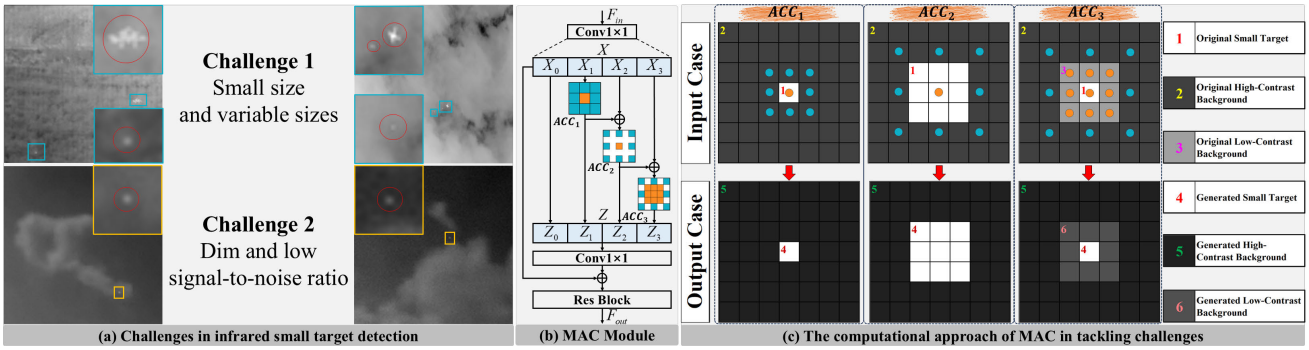

Fig. 3. (a) Two challenges existed in infrared small-target detection. (b) Internal structure of our MAC. ACC_1, ACC_2, and ACC_3 are three different atrous contrast convolution (ACC) kernels. The result of each convolution kernel is obtained by subtracting the average value of the pixels in the blue grids from the average value of the pixels in the yellow grid. (c) Computational approach of MAC in tackling challenges.

Fig. 3(b) illustrates the internal structure of our MAC module, which comprises four branches: one direct connection branch and three parallel atrous contrast branches. Specifically, given the input feature map F_{in} (for simplicity, we omit stage indices $i \in \{1, 2, 3, 4, 5\}$ for MAC inputs F_{in}^i and outputs F_{out}^i shown in Fig. 2), we first expand the channel dimension and produce the feature map X, using the following equation:

$$X = \mathrm{Conv}_{1\times1}(F_{\text{in}}) \tag{1}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ refers to the 1×1 convolutional block and aims to realize the channel dimension expansion. Subsequently, the feature map X is evenly divided into four groups along the channel dimension, denoted as $X = [X_0, X_1, X_2, X_3]$. X_0 serves as the direct connection and is output immediately, labeled as Z_0. X_1, X_2, and X_3 are in parallel fed into three different ACC branches, namely ACC_1, ACC_2, and ACC_3, respectively, producing the contrastive feature with a different receptive field, denoted as Z_1, Z_2, and Z_3. These operations are formulated as the following equation:

$$Z_i = \begin{cases} X_i, & i = 0 \\ \mathrm{ACC}_1(X_i), & i = 1 \\ \mathrm{ACC}_2(X_i + Z_{i-1}), & i = 2 \\ \mathrm{ACC}_3(X_i + Z_{i-1}), & i = 3 \end{cases} \tag{2}$$

where $\mathrm{ACC}_1(\cdot)$, $\mathrm{ACC}_2(\cdot)$, and $\mathrm{ACC}_3(\cdot)$ represent three parallel MAC convolution kernels. The three convolution kernels vary in size and atrous rates: $\mathrm{ACC}_1(s = 3, d = 1, c = 1)$, $\mathrm{ACC}_2(s = 5, d = 2, c = 1)$, and $\mathrm{ACC}_3(s = 5, d = 2, c = 3)$. Here, "s" denotes the size of the kernel, "d" represents the dilation rate, and "c" indicates the size of the central yellow region. These kernels compute contrast by subtracting the mean value of the surrounding blue region from the mean value of the central yellow region, enabling them to capture a wider range of contextual information. The calculation method is formulated as the following equation:

$$\mathrm{ACC}_i(X_i) = \frac{1}{n}\sum_{(u,v)\in Y} X_i(u, v) - \frac{1}{m}\sum_{(u,v)\in B} X_i(u, v) \tag{3}$$

where (u, v) represents the pixel coordinates in the feature map X_i. Y refers to the set of central pixels marked yellow and B refers to the set of surrounding pixels marked blue in one contrast convolution kernel. n and m represent the total number of central or surrounding pixels with one contrast kernel. Then, all the features $Z = [Z_0, Z_1, Z_2, Z_3]$ are concatenated along the channel dimension and passed through a 1×1 convolution for information mixing. A residual block is conducted on Z and X to obtain a contrast-enhanced feature map, using the following equation:

$$F_{\text{out}} = \mathrm{Res\_Block}(\mathrm{Conv}_{1\times1}(Z) + X) \tag{4}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ refers to the 1×1 convolutional block. The $\mathrm{Res\_Block}(\cdot)$ consists of a $\mathrm{Conv}_{3\times3}(\cdot)$, channel attention $\mathrm{CA}(\cdot)$, and spatial attention $\mathrm{SA}(\cdot)$.

To provide an intuitive understanding of the MAC module, the computational approach of MAC in tackling challenges is illustrated in Fig. 3(c). Each pixel block of different colors represents different targets and backgrounds, as marked in the rightmost column of Fig. 3(c). The three input cases at the top are processed by three different convolution kernels to generate the corresponding output cases at the bottom, respectively. In each set of input–output cases, the input pixel blocks (marked 1, 2, and 3) are processed to generate the corresponding output pixel blocks (marked 4, 5, and 6). For ACC_1, a simple contrast convolution kernel is used, which enhances the contrast between small targets with very few pixels and the background, verifying its ability to perceive extremely small targets. For ACC_2, the contrast convolution kernel is dilated, which enhances the contrast between small targets with relatively more pixels and the background, verifying its ability to perceive slightly larger targets. For ACC_3, the central pixels of the kernel are expanded to a 3×3 region, enabling effective detection of small targets in low signal-to-noise ratio areas. ACC_3 enhances the contrast between small targets and the background, while suppressing the low-contrast background blocks (marked 3, suppressed to marked 6) surrounding the target, verifying its ability to perceive small targets in low signal-to-noise ratio conditions.

C. DHPF Block in the Frequency Domain

Many IRIs contain extensive low-frequency background regions, such as the highlighted road surface area at the bottom of Fig. 4(b), along with sparse high-frequency regions, including small targets and edges between the highlighted road surface and the black background. The traditional CNN methods in the spatial domain operate akin to low-pass filters, adept

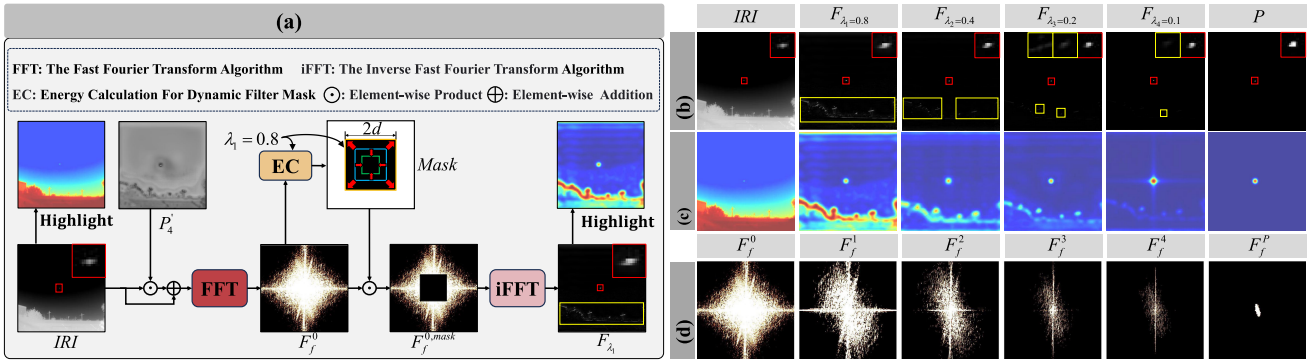

Fig. 4. (a) Internal structure of the first DHPF stage, with the remaining stages having different λ values, while the computational process remains the same. (b) Original input image IRI, the enhanced feature maps (F_{λ_1}, F_{λ_2}, F_{λ_3}, and F_{λ_4}) of four DHPF stages, and the final prediction map P. (c) Color-highlighted version of (b). (d) After FFT processing, F_f^0 refers to the frequency feature map of the IRI. F_f^1, F_f^2, F_f^3, and F_f^4 refer to the frequency feature maps enhanced by four-stage DHPFs, and F_f^P refers to the frequency feature map of the final prediction map.

at eliminating high-frequency information through kernel-based aggregation, thereby failing to meet our need to preserve high-frequency details while eliminating low-frequency ones. Additionally, existing frequency-domain learning methods often rely on fixed thresholds to eliminate low-frequency information, limiting their ability to handle IRSTD scenarios with diverse frequency energy distributions. To tackle this problem, we propose a DHPF module. In the frequency-domain branch (Fig. 2), we integrate customized DHPFs at different stages to progressively remove low-frequency components from the image.

Fig. 4(a) depicts the internal structure of the first DHPF stage. Upon receiving the input IRI, it undergoes initial filtering using the prediction map P_4' generated by the spatial-domain decoder, effectively reducing background interference and yielding a targets-enhanced image. Then, the target-enhanced image is transformed into a frequency map via FFT. It is formulated in the following equation:

$$F_f^0 = \mathrm{FFT}\left(\mathrm{IRI} \odot P_4' + \mathrm{IRI}\right) \tag{5}$$

where $\mathrm{IRI} \in \mathbb{R}^{H\times W\times 1}$ and $P_4' \in \mathbb{R}^{H\times W\times 1}$ represent the original image and the prediction map generated by the corresponding decoder stage, respectively. $\odot$ means the element-wise production. $F_f^0 \in \mathbb{R}^{H\times W\times 1}$ refers to the frequency feature map, and $\mathrm{FFT}(\cdot)$ refers to the FFT.

The frequency feature map F_f^0 illustrated in Fig. 4(a) demonstrates that as one approaches the center, the frequency decreases and the amplitude increases, whereas as one moves away from the center, the frequency increases and the amplitude decreases. This indicates that the IRI contains abundant low-frequency background information. Since both the small target area and the edge area belong to high-frequency information, to effectively eliminate the low-frequency background and high-frequency edge background information, we propose a method of progressively filtering out a certain proportion of low-frequency information. Specifically, we first calculate the energy of the frequency feature map F_f^0, using the following equation:

$$\mathrm{EC} = \sum_{u=1}^{H}\sum_{v=1}^{W}\left|F_f^0(u,v)\right|^2 \tag{6}$$

where (u, v) represents the pixel coordinates in the frequency feature map. $F_f^0(u, v)$ is the amplitude value at pixel (u, v), and EC represents the total frequency energy of the image.

Then, we calculate the filter mask based on the preset energy removed ratio. The formula in the 1st DHPF is as follows:

$$E_{\mathrm{removed}} = \sum_{u=u_0-d}^{u_0+d}\sum_{v=v_0-d}^{v_0+d}\left|F_f^0(u,v)\right|^2 \le \lambda_1 \times \mathrm{EC} \tag{7}$$

where E_{removed} represents the low-frequency energy to be suppressed at the DHPF stage. (u_0, v_0) indicates the central pixel coordinates of the feature map. λ_1 refers to the preset energy removed ratio. d denotes the maximum radius value that satisfies (7). Therefore, we can obtain the dynamic filtering mask as follows:

$$\mathrm{Mask}(u,v) = \begin{cases} 0, & \text{if } |u-u_0| \le d,\ |v-v_0| \le d \\ 1, & \text{others} \end{cases} \tag{8}$$

where $\mathrm{Mask} \in \mathbb{R}^{H\times W\times 1}$ represents the dynamic filtering mask. Since the initial IRI encompasses abundant low-frequency information, and as the decoder transitions to shallower layers, its predicted maps exhibit a decreasing amount of low-frequency background. Therefore, we gradually decrease the energy filtering ratios and preset them as $\lambda = [\lambda_1, \lambda_2, \lambda_3, \lambda_4] = [0.8, 0.4, 0.2, 0.1]$.

We can obtain the target-enhanced feature map from the DHPF module, using the following equation:

$$F_{\lambda_1} = \mathrm{iFFT}\left(\mathrm{Mask} \odot F_f^0\right) \tag{9}$$

where $\mathrm{iFFT}(\cdot)$ represents the inverse FFT and F_{λ_1} refers to the output of the 1st DHPF module. The visualized feature maps in the frequency domain are shown in Fig. 4(d), where it can be observed that as λ decreases in different DHPF modules, the low-frequency information decreases. The spatial-domain predicted maps shown in the second–fifth columns of Fig. 4(b) and (c) demonstrate that, as λ decreases, the large areas of low-frequency background information and high-frequency edge background areas are gradually suppressed.

In the end, the final predicted map P can be obtained by fusing the predicted map P_0 in the spatial domain and the feature map F_{λ_4} generated from the 4th DHPF stage in the

frequency domain. It is formulated in the following equation:

$$P = P_0 \odot \mathrm{sig}(F_{\lambda_4}) + P_0 \tag{10}$$

where $\mathrm{sig}(\cdot)$ is the sigmoid function. P_0 is obtained by concatenating all the intermediate decoder stage prediction maps and using $\mathrm{Conv}_{1\times1}$ to reduce the number of channels. As shown in Fig. 4(b) and (c), the last column labeled "P" represents the final predicted map P of our model, which indicates that our model can effectively suppress the background areas while preserving the small target areas.

D. Loss Function

We employed the scale and location sensitive (SLS) loss [6] $\mathcal{L}_{\mathrm{SLS}}$ to supervise our HDNet, by minimizing the difference between predicted maps and ground truth, using the following equation:

$$\mathcal{L}_{\mathrm{SLS}} = \mathcal{L}_S + \mathcal{L}_L \tag{11}$$

where $\mathcal{L}_S$ and $\mathcal{L}_L$ refer to the scale-sensitive loss and location-sensitive loss, respectively.

$\mathcal{L}_S$ is defined as

$$\begin{aligned} \mathcal{L}_S &= 1 - w\frac{|P \cap G|}{|P \cup G|} \\ \text{s.t. } w &= \frac{\min(|P|,|G|) + \mathrm{Var}(|P|,|G|)}{\max(|P|,|G|) + \mathrm{Var}(|P|,|G|)} \end{aligned} \tag{12}$$

where P and G represent the prediction map and ground-truth map, respectively. $|\cdot|$ represents the count of the pixel set. $\mathrm{Var}(\cdot,\cdot)$ computes the variance of given scalars.

$\mathcal{L}_L$ is defined as

$$d_p = \sqrt{x_p^2 + y_p^2}, \quad \theta_p = \arctan\left(\frac{y_p}{x_p}\right) \tag{13}$$

$$d_{\mathrm{gt}} = \sqrt{x_{\mathrm{gt}}^2 + y_{\mathrm{gt}}^2}, \quad \theta_{\mathrm{gt}} = \arctan\left(\frac{y_{\mathrm{gt}}}{x_{\mathrm{gt}}}\right) \tag{14}$$

$$\mathcal{L}_L = \left(1 - \frac{\min(d_p, d_{\mathrm{gt}})}{\max(d_p, d_{\mathrm{gt}})}\right) + \frac{4}{\pi^2}(\theta_p - \theta_{\mathrm{gt}})^2 \tag{15}$$

where $\mathbf{c}_p = (x_p, y_p)$ and $\mathbf{c}_{\mathrm{gt}} = (x_{\mathrm{gt}}, y_{\mathrm{gt}})$ are the central points of the predicted target pixel set in P and the ground-truth target pixel set in G, respectively. They are calculated by averaging the coordinates of all pixels in each pixel set. Then, we convert the coordinates of these center points into the polar coordinate system. Then, (d_p, θ_p) and $(d_{\mathrm{gt}}, \theta_{\mathrm{gt}})$ refer to the distance and angle of $\mathbf{c}_p$ and $\mathbf{c}_{\mathrm{gt}}$, respectively.

We adopted $\mathcal{L}_{\mathrm{SLS}}$ to supervise all five predicted maps generated from different stages. It is formulated in the following equation:

$$\mathcal{L} = \frac{1}{5}\left(\sum_{i=1}^{4} \mathcal{L}_{\mathrm{SLS}}\left(P_i, \Downarrow\left(G, 2^{i-1}\right)\right) + \mathcal{L}_{\mathrm{SLS}}(P, G)\right) \tag{16}$$

where $\Downarrow(\cdot,\cdot)$ is the operation that spatially downsamples the first argument with the second argument as the factor. P_i ($i \in [1, 2, 3, 4]$) refers to the predicted maps obtained from four decoder stages and P refers to the final prediction map. G is the ground-truth map.

IV. Experiment

In Section IV-A, we introduce the experimental setup, including publicly available datasets, implementation details, and evaluation metrics. Section IV-B presents a comprehensive comparison of our HDNet with 26 advanced models, confirming its superiority. Extensive ablation studies in Section IV-C validate the effectiveness of our network components. Section IV-D compares the model complexity of our HDNet with several SOTA methods, demonstrating its superior balance between efficiency and performance.

A. Experimental Settings

1) Datasets: We conducted experiments on three commonly used datasets, including IRSTD-1k [23], NUAA-SIRST [25], and NUDT-SIRST [24], which contain 1001, 427, and 1327 IRIs, respectively. Following existing works [23], [24], [25], the images in NUAA-SIRST and NUDT-SIRST are divided equally into training sets and testing sets, while the images in IRSTD-1k are divided into training sets and testing sets with a 4:1 ratio.

2) Implementation Details: Our HDNet is constructed under the PyTorch framework and runs on a single RTX 3090 GPU. Then, we augment the training data by random flipping and rotation. Following existing works, we resize the input image to 256×256. The model is trained for 800 epochs with a batch size of 4, utilizing the AdaGrad optimizer. The initial learning rate is set to 0.05 and is adjusted using the CosineAnnealingLR Learning Rate Scheduler with a period of 50 epochs, allowing the learning rate to decline smoothly to 0.045 over each period.

3) Evaluation Metrics: We adopted several commonly used metrics to evaluate our proposed HDNet and existing methods: intersection over union (IoU) and false alarm rate (F_a) as the pixel-level evaluation metrics, and probability of detection (P_d) as the object-level evaluation metric. We also compared our method with the most advanced methods using the receiver operating characteristic (ROC) curves.

The IOU is defined as

$$\mathrm{IoU} = \frac{A_i}{A_u} = \frac{\sum_{j=1}^{N} \mathrm{TP}[j]}{\sum_{j=1}^{N}\left(T\left[j\right] + P\left[j\right] - \mathrm{TP}\left[j\right]\right)} \tag{17}$$

where A_i and A_u represent the sizes of intersection area and union area, respectively. N is the total number of pixels, $\mathrm{TP}[\cdot]$ denotes the number of predicted true positive pixels, $T[\cdot]$ represents the number of true positive pixels in ground truth, and $P[\cdot]$ represents the number of predicted positive pixels.

The probability of detection is defined as

$$P_d = \frac{N_p}{N_{\mathrm{all}}} \tag{18}$$

where N_p is the count of accurately predicted targets and N_{all} denotes the total number of all targets.

The false alarm rate is defined as

$$F_a = \frac{P_f}{P_{\mathrm{all}}} \tag{19}$$

where P_f is the number of falsely predicted target pixels and P_{all} is the total number of pixels in the image.

TABLE I

QUANTITATIVE COMPARISONS BETWEEN OUR HDNET AND 26 SOTA METHODS IN TERMS OF IOU (%), P_d (%), AND F_a (10^{-6}). THE OPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD, AND THE SECOND-BEST SCORES ARE UNDERLINED. THE SYMBOL "-" DENOTES NO DATA AVAILABLE. THE METHODS ARE CATEGORIZED AS: TRAD-F (TRADITIONAL FILTER-BASED METHODS), TRAD-C (TRADITIONAL LOCAL CONTRAST-BASED METHODS), TRAD-L (TRADITIONAL LOW-RANK BASED METHODS), CNN (DEEP CONVOLUTION METHODS), HYBRID-T (HYBRID CNN AND TRANSFORMER METHODS), AND HYBRID-D (HYBRID DOMAIN METHOD). ↑ AND ↓ SIGNIFIES BETTER PERFORMANCE WITH HIGHER VALUES AND LOWER VALUES, RESPECTIVELY

Method	Type	Size	Year	IRSTD-1k			NUAA-SIRST			NUDT-SIRST		
				$IoU\uparrow$	$P_d\uparrow$	$F_a\downarrow$	$IoU\uparrow$	$P_d\uparrow$	$F_a\downarrow$	$IoU\uparrow$	$P_d\uparrow$	$F_a\downarrow$
Max-Median [11]	Trad-F	256	1999	6.70	65.21	59.73	6.02	84.34	774.3	4.20	58.41	36.89
Top-Hat [43]	Trad-F	256	2010	10.06	75.11	1432	1.51	79.74	16456	20.72	78.41	166.7
IPI [19]	Trad-L	256	2013	27.92	81.37	16.18	1.09	87.05	30467	17.76	74.49	41.23
RIPT [18]	Trad-L	256	2017	14.11	77.55	28.31	16.79	69.76	59.33	29.44	91.85	344.3
NRAM [44]	Trad-L	256	2018	15.25	70.68	16.93	15.25	70.68	16.93	6.93	56.40	19.27
PSTNN [21]	Trad-L	256	2019	24.57	71.99	35.26	30.30	72.80	48.99	14.85	66.13	44.17
MSLSTIPT [20]	Trad-L	256	2021	11.43	79.03	1524	1.08	0.052	8.18	8.34	47.40	888.1
TLLCM [14]	Trad-C	256	2020	3.31	77.39	6738	4.24	88.37	6243	2.18	62.01	1608
WSLCM [15]	Trad-C	256	2021	3.45	72.44	6619	6.39	88.74	4462	2.28	56.82	1309
MDvsFA [30]	CNN	256	2019	37.34	83.71	88.52	60.30	89.35	56.35	35.86	85.22	95.37
ALCNet [31]	CNN	256	2021	65.68	89.25	27.71	73.74	97.25	26.79	72.89	96.19	30.40
ACMNet [25]	CNN	256	2021	60.33	93.27	68.49	69.44	92.02	22.71	64.86	96.72	28.59
ISNet [23]	CNN	256	2022	61.85	90.24	31.56	70.49	95.06	67.98	81.24	97.78	6.34
DNANet [24]	CNN	256	2022	65.71	91.84	17.61	77.76	96.33	2.31	79.98	96.93	12.78
RKformer [45]	Hybrid-T	512	2022	64.12	93.27	18.65	77.24	99.11	1.58	-	-	-
ISTDU-Net [46]	CNN	512	2022	65.01	93.94	26.44	75.93	96.20	38.90	91.76	98.52	3.77
UIU-Net [32]	CNN	256	2023	68.69	91.25	13.48	77.53	92.40	9.33	75.91	96.83	18.61
RDIAN [47]	CNN	256	2023	59.94	87.21	33.31	70.74	95.06	48.16	82.42	96.72	14.85
MTUNet [48]	Hybrid-T	256	2023	64.09	90.48	12.15	74.85	99.08	7.09	77.98	96.08	17.51
RPCANet [49]	CNN	256	2024	63.21	88.31	4.39	65.08	93.58	10.85	89.31	97.14	2.87
MSHNet [6]	CNN	256	2024	67.16	93.88	15.03	73.5	97.25	31.05	80.55	97.99	11.77
GCI-Net [50]	CNN	512	2024	67.75	93.89	12.84	78.81	99.34	2.11	-	-	-
SCTransNet [5]	Hybrid-T	256	2024	68.03	93.27	10.74	77.5	96.95	13.92	**94.09**	**98.62**	4.29
PBT [4]	Hybrid-T	256	2024	68.49	92.52	8.88	78.39	99.08	2.13	83.89	97.23	4.23
L^2SKNet [51]	CNN	256	2025	67.81	90.24	17.46	73.43	98.17	20.82	93.58	97.57	5.33
MMLNet [52]	CNN	256	2025	67.21	94.28	14.00	78.71	98.88	25.71	81.81	98.43	11.77
HDNet(Ours)	Hybrid-D	256		**70.26**	**94.56**	**4.33**	**79.17**	**100**	**0.53**	85.17	98.52	**2.78**

The ROC curves are plotted based on different true-positive rates (TPR) and false-positive rates (FPR) under different thresholds. The TPR is defined as

$$\mathrm{TPR} = \frac{\sum_{j=1}^{N} \mathrm{TP}[j]}{\sum_{j=1}^{N} \left(\mathrm{TP}[j] + \mathrm{FN}[j]\right)} \tag{20}$$

where N and TP[·] are same to (17). FN[·] denotes the number of predicted false-negative pixels.

The FPR is defined as

$$\mathrm{FPR} = \frac{\sum_{j=1}^{N} \mathrm{FP}[j]}{\sum_{j=1}^{N} \left(\mathrm{TN}[j] + \mathrm{FP}[j]\right)} \tag{21}$$

where N is same to (17). FP[·] denotes the number of predicted false-positive pixels and TN[·] denotes the number of predicted true-negative pixels.

B. Comparison With SOTA Methods

1) Comparison Methods: We conducted a comprehensive comparison of our method with 26 SOTA methods on three challenging datasets. For traditional methods, including nine well-established methods (Max–Median [11], Top-Hat [43], IPI [19], RIPT [18], NRAM [44], PSTNN [21], MSLSTIPT [20], TLLCM [14], and WSLCM [15]). For representative deep learning-based IRSTD methods, we selected MDvsFA [30], ALCNet [31], ACMNet [25], ISNet [23], DNANet [24], RKformer [45], ISTDU-Net [46], UIUNet [32], RDIAN [47], MTUNet [48], RPCANet [49], MSHNet [6], GCI-Net [50], SCTransNet [5], PBT [4], L^2SKNet [51], and MMLNet [52]. For a fair comparison, all the predicted maps are collected from publicly available results provided by the authors or computed by running the source code released by the authors.

2) Quantitative Comparison: Table I presents a primary quantitative comparison between our HDNet and 26 advanced methods on three challenging benchmark datasets. It can be observed that our proposed HDNet consistently outperforms these advanced methods on the IRSTD-1k and NUAA-SIRST datasets and also improves the performance on the NUDT-SIRST dataset with the lowest false alarm rate F_a. Among all the nine metrics reported in Table I, our HDNet generally achieves seven best performances, one second-best performance, and one fifth-best performance. Especially, our HDNet achieves a target-level P_d accuracy of 100% on the NUAA-SIRST dataset. Although our HDNet obtains a slightly lower performance than SCTransNet in terms of IoU

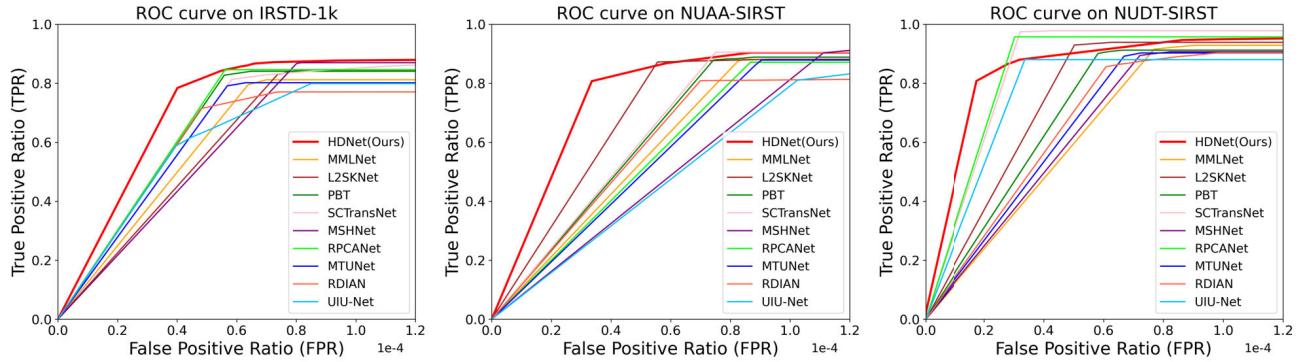

Fig. 5. ROC curves of different methods on the IRSTD-1k, NUAA-SIRST, and NUDT-SIRST datasets. The closer the curves to the top-left corner, the better performance.

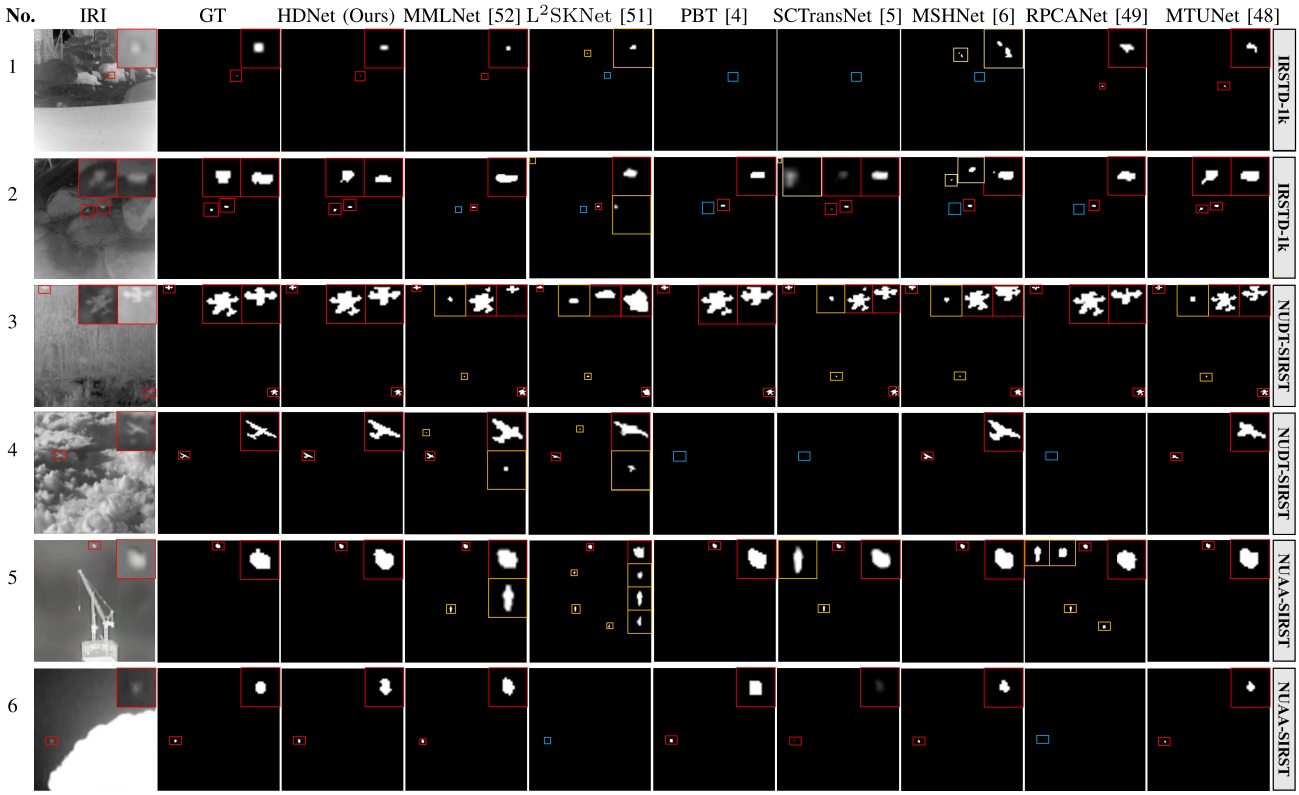

Fig. 6. Qualitative comparisons between our HDNet and seven most recent advanced methods. The correctly detected targets, false alarms, and missed targets are framed by red, yellow, and blue bounding boxes, respectively. For better visualization, a close-up view of the target is shown in image corners.

and P_d metrics on the NUDT-SIRST dataset, our method consistently outperforms it on the other two datasets. It is also worth noting that SCTransNet is a hybrid method that integrates CNNs and Transformers [53]. As demonstrated in Table VI, the SCTransNet has a significantly higher number of parameters and computational complexity compared to our model. Additionally, compared to another latest hybrid Transformer-based method (PBT [4]), our HDNet gains performance improvements of 1.77%, 0.78%, and 1.28% in terms of IoU on the IRSTD-1k, NUAA-SIRST, and NUDT-SIRST datasets, and 2.04%, 0.92%, and 1.29% in terms of P_d on the IRSTD-1k, NUAA-SIRST, and NUDT-SIRST datasets. Compared to the latest MMLNet [52] method, our method achieves a higher performance in terms of all metrics on the three datasets. For L2SKNet [51], our method achieves higher performance on IRSTD-1k and NUAA-SIRST datasets and falls behind L2SKNet in terms of IoU metric on the NUDT-SIRST dataset. It is attributed that L2SKNet may benefit from its combining various CNNs and attention methods, guiding the network to capture salient features of infrared targets. In terms of the F_a metric, our HDNet achieves the lowest false alarm rate across all three datasets, clearly demonstrating the effectiveness of the proposed MAC and DHPF modules in identifying small targets and suppressing the complex background.

We also presented the ROC curves of the advanced methods published in the latest three years on the three datasets, as shown in Fig. 5. It can be observed that the HDNet performs the best on the IRSTD-1k and NUAA-SIRST datasets while achieving a higher TPR at a lower FPR on the NUDT-SIRST dataset, fully demonstrating HDNet's competitiveness compared to other advanced methods.

3) Qualitative Comparison: As depicted in Fig. 6, these visual examples are employed to conduct a qualitative analysis of our HDNet in comparison with seven of the most recent advanced methods, namely MMLNet [52], L^{2}SKNet [51], PBT [4], SCTransNet [5], MSHNet [6], RPCANet [49], and MTUNet [48]. Upon examining the original IRIs presented in the first column of Fig. 6, it becomes apparent that infrared small targets typically exhibit two particularly challenging characteristics. First, they are small and exhibit variable sizes (as evidenced in the first, third, and fourth rows). A notable example is the small target in the first row, which occupies just a handful of pixels, whereas the small targets in the third and fourth rows possess irregular shapes and may span several 100 pixels. This presents a significant challenge for existing models, as they struggle to effectively perceive small targets of varying sizes, often leading to missed targets or false alarms. Second, another distinctive characteristic is that some small targets appear faint and blend seamlessly into the background or are interfered with bright background noise, making it difficult to distinguish them from their immediate surroundings (as illustrated in the first, second, fifth, and sixth rows). Specifically, the small targets in the first, second, and sixth rows exhibit similarities in appearance to the background, while the 5th row demonstrates instances of bright background noise. We also conducted a comparative analysis of visual examples against two advanced methods, MMLNet [52] and L2SKNet [51]. As shown in Fig. 6, our method excels in accurately segmenting multiscale targets and suppressing background noise. Both MMLNet and L2SKNet exhibit varying degrees of false detections and missed detections (rows 2–5, columns 4–5). Existing methods face difficulties in differentiating the targets from the background, often mistakenly identifying the background noise as the foreground, thereby resulting in an elevated false alarm rate. However, our method excels in detecting these small targets. This can be attributed to the efficacy of our proposed MAC module, which possesses the capability to perceive small targets of varying sizes, combined with the DHPF module, which adeptly suppresses both redundant low-frequency backgrounds and high-frequency background noise.

C. Ablation Experiments

In this section, we conducted comprehensive ablation studies to validate the effectiveness of our proposed modules on the IRSTD-1k and NUDT-SIRST datasets. Specifically, our investigation includes ablation experiments between different modules and ablation experiments in each module.

1) Ablation Experiments Between Different Modules: To validate the effectiveness of our proposed MAC and DHPF components, we designed four network variants: 1) baseline (the basic model equipped with only the plain encoder–decoder structure); 2) baseline + MAC (the variant model equipped with the MAC component in the encoder); 3) baseline + DHPF (the variant model equipped with the DHPF component in frequency branch); and 4) baseline + MAC + DHPF (the full HDNet). The quantitative results are reported in Table II. On the IRSTD-1k dataset, we observed

TABLE II

QUANTITATIVE RESULTS OF THE ABLATION EXPERIMENTS BETWEEN DIFFERENT MODULES ON THE IRSTD-1K AND NUDT-SIRST DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

No.	MAC	DHPF	IRSTD-1k			NUDT-SIRST		
			$IoU\uparrow$	$P_d\uparrow$	$F_a\downarrow$	$IoU\uparrow$	$P_d\uparrow$	$F_a\downarrow$
1			62.08	89.80	34.54	77.36	95.45	17.37
2	✓		68.12	94.22	11.69	82.93	97.67	6.85
3		✓	67.73	93.54	12.22	82.92	97.57	13.17
4	✓	✓	**70.26**	**94.56**	**4.33**	**85.17**	**98.52**	**2.78**

TABLE III

QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS INSIDE THE MAC MODULE. THE BEST RESULTS ARE MARKED IN BOLD

No.	Model Variants	IRSTD-1k			NUDT-SIRST		
		$IoU\uparrow$	$P_d\uparrow$	$F_a\downarrow$	$IoU\uparrow$	$P_d\uparrow$	$F_a\downarrow$
1	wo atrous	69.37	93.54	7.97	**85.19**	97.99	4.67
2	wo contrast	68.12	92.52	11.08	82.88	97.88	10.25
3	all_ACC_1	66.16	94.22	21.79	82.10	97.78	9.84
4	all_ACC_2	68.07	93.54	8.81	82.90	98.20	6.39
5	all_ACC_3	67.75	92.18	7.90	81.75	98.10	9.93
6	Ours	**70.26**	**94.56**	**4.33**	85.17	**98.52**	**2.78**

that the "baseline" (Row No. 1) achieves 0.6208 in terms of IoU and 0.8980 in terms of P_d. The proposed MAC component ("baseline + MAC," Row No. 2) improves these metrics by 6.04% and 4.42%, respectively. Additionally, the DHPF module ("baseline + DHPF," Row No. 3) increases these metrics by 5.65% and 3.74%, respectively. Both MAC and DHPF components can effectively reduce F_a. With the equipment of both MAC and DHPF, the complete HDNet ("baseline + MAC + DHPF," Row No. 4) achieves the best performance and significantly outperforms the "baseline" model, achieving improvements of 8.18%–4.76% in terms of IoU and P_d, respectively. A similar trend can also be observed in the NUDT-SIRST dataset. All these quantitative results demonstrate the effectiveness of our MAC and DHPF, which contribute to the perception of targets with small and variable size, and suppression of low-frequency background.

In addition, the visual examples shown in Fig. 7(a) demonstrate the benefits of our proposed components. The 'baseline' variant produces the worst segmentation results, with severe false alarms and missed targets in multiscale, dim small-target detection. The "baseline + DHPF" (without the MAC module) variant, as shown in the first and second rows, fifth column of Fig. 7(a), exhibits significant missed detection of the small dim target. This highlights the crucial role of the MAC module in improving the model's ability to perceive variable size and small dim targets. The "baseline + MAC" (without the DHPF module), as shown in the second row, fourth column of Fig. 7(a), the model displays notable false alarms in small dim target detection, whereas our "baseline + MAC + DHPF" detects targets more comprehensively. This emphasizes the importance of the DHPF module in enhancing target information and suppressing background noise.

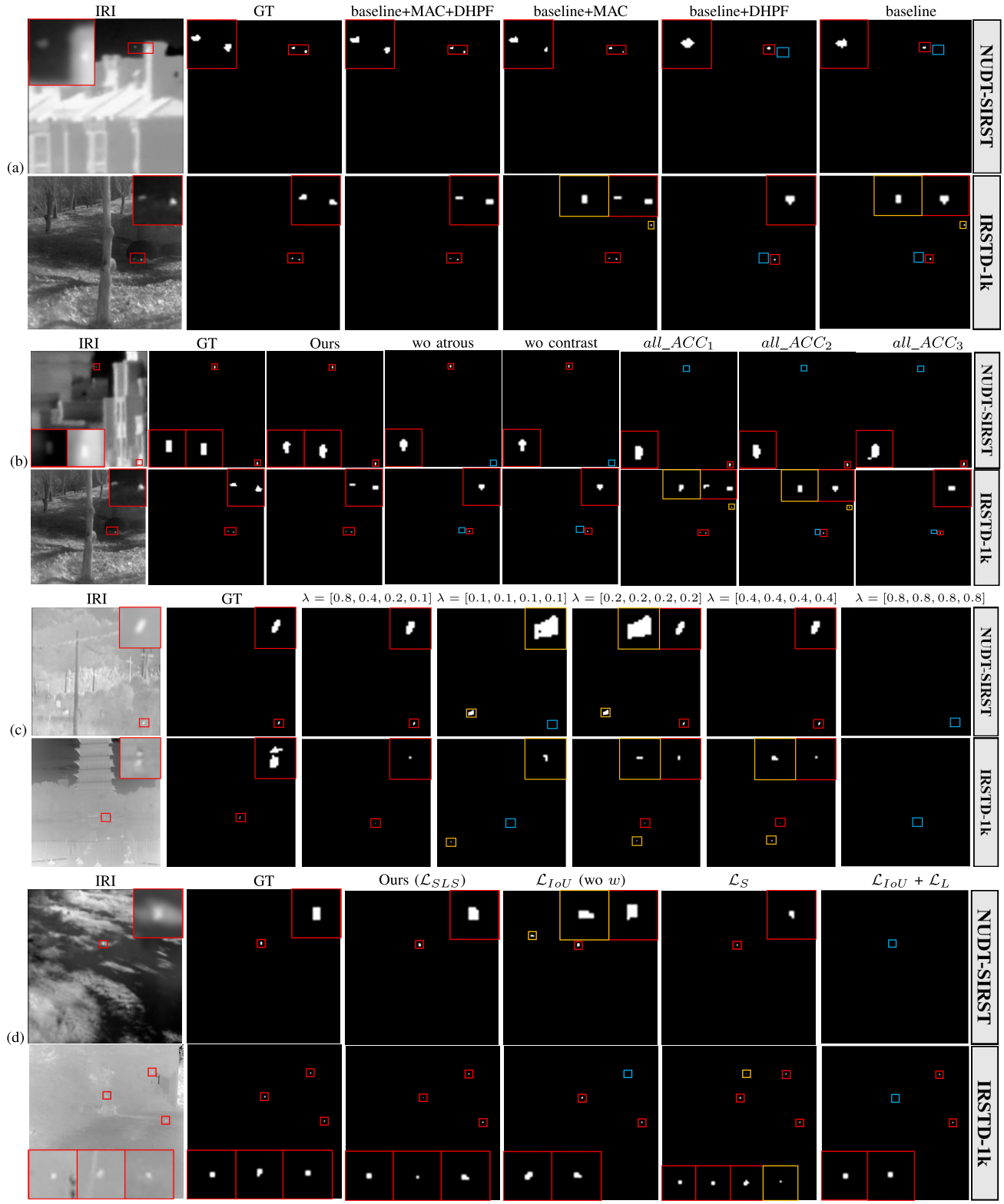

Fig. 7. Visual examples of ablation experiments on the IRSTD-1k and NUDT-SIRST dataset. (a) Between different modules. (b) Inside the MAC module. (c) Inside the HDPF module. (d) Inside the SLS loss. The correctly detected targets, false alarms, and missed targets are framed by red, yellow, and blue bounding boxes, respectively. For better visualization, a close-up view of the target is shown in image corners.

2) Ablation Experiments Inside the MAC Module: To validate the effectiveness of MAC convolutions in the MAC module, we provided five additional MAC variants: 1) "wo atrous" (three filters without atrous concept, the atrous portion is filled with blue); 2) "wo contrast" (three contrast kernels are replaced with plain convolution kernels); 3) "all_ACC$_1$" (all three kernels are configured with ACC$_1$ kernel); 4) "all_ACC$_2$" (all three kernels are configured with ACC$_2$ kernel); and 5) "all_ACC$_3$" (all three kernels are configured with ACC$_3$ kernel). The quantitative results are reported in Table III, where

the most performance metrics of all variants decrease compared with our full model. For the "wo atrous," except for a slightly better IoU on NUDT-SIRST than our full model, other indicators all declined. This validates the impact of our atrous contrast idea for perceiving small targets with variable sizes. For the "wo contrast," all the performance indicators declined compared to our full model, which demonstrates the effectiveness of contrast computing in identifying the small targets. For the "all_ACC$_1$," "all_ACC$_2$," and "all_ACC$_3$," all the performance indicators declined compared to our full model. This verifies the effectiveness of our MAC in perceiving targets of small and variable size. We concluded that, by designing three different scales of atrous convolutional kernels, the MAC module can significantly expand the model's receptive field without changing resolution and increase the contrast between target and background with a low signal-to-noise ratio, boosting the target extraction of various sizes from complex backgrounds.

The visual examples shown in Fig. 7(b) further confirm the effectiveness of our MAC module. The five model variants perform poorly in detecting small and dim infrared targets in complex backgrounds, exhibiting varying degrees of false alarms and missed targets. "wo atrous" (fourth column) fails to fully recognize multiscale small targets. "wo contrast" (fifth column) lacks accurate detection of dim targets. "all_ACC$_1$," "all_ACC$_2$," and "all_ACC$_3$," due to using a single scale contrast convolution kernel, exhibit missed targets and false alarms. Our final full method could accurately detect all small targets. This also indicates that our MAC module effectively combines the advantages of atrous and contrast convolutions, accurately extracting multiscale small and dim targets in complex backgrounds.

3) Ablation Experiments Inside the DHPF Module: To validate the effectiveness of the energy filtering ratios in the DHPF module, we provided four additional DHPF variants: 1) $\lambda = [0.1, 0.1, 0.1, 0.1]$ (the energy filtering ratio sets to 0.1 for all four stages); 2) $\lambda = [0.2, 0.2, 0.2, 0.2]$ (the energy filtering ratio sets to 0.2 for all four stages); 3) $\lambda = [0.4, 0.4, 0.4, 0.4]$ (the energy filtering ratio sets to 0.4 for all four stages); and 4) $\lambda = [0.8, 0.8, 0.8, 0.8]$ (the energy filtering ratio sets to 0.8 for all four stages). The quantitative results are reported in Table IV, where the performance metrics of all variants decrease compared with our full model. For "$\lambda = [0.1, 0.1, 0.1, 0.1]$" and "$\lambda = [0.2, 0.2, 0.2, 0.2]$," as the energy filtering ratio increases, all performance metrics improve, although none reach the SOTA levels. This partially demonstrates the effectiveness of the DHPF module in suppressing background information and gradually enhancing foreground information. For "$\lambda = [0.4, 0.4, 0.4, 0.4]$," when the energy filtering ratio is uniformly increased, the model misses some target information, leading to a decrease in P_d accuracy (Row Nos. 2–3). For "$\lambda = [0.8, 0.8, 0.8, 0.8]$," further increasing the energy filtering ratio leads to the removal of some target information, causing a decrease in most performance metrics. This validates the effectiveness of our DHPF in progressively filtering out slowly changing, rich low-frequency backgrounds. We concluded that, by setting different levels of energy filtering ratios in the frequency domain, the model can gradually filter out slowly changing, rich low-frequency backgrounds in four stages, without interfering with the target information. This progressive filtering of low-frequency information enhances the representation of foreground targets.

TABLE IV
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS INSIDE THE DHPF MODULE. THE BEST RESULTS ARE MARKED IN BOLD

No.	Model Variants	IRSTD-1k			NUDT-SIRST		
		$IoU \uparrow$	$P_d \uparrow$	$F_a \downarrow$	$IoU \uparrow$	$P_d \uparrow$	$F_a \downarrow$
1	$\lambda = [0.1, 0.1, 0.1, 0.1]$	66.98	93.20	14.35	81.93	97.14	12.36
2	$\lambda = [0.2, 0.2, 0.2, 0.2]$	67.70	93.54	12.68	82.06	98.41	11.61
3	$\lambda = [0.4, 0.4, 0.4, 0.4]$	68.17	92.18	7.21	83.67	97.78	7.19
4	$\lambda = [0.8, 0.8, 0.8, 0.8]$	67.94	92.86	12.15	83.96	97.67	10.46
5	$\lambda = [0.8, 0.4, 0.2, 0.1]$	**70.26**	**94.56**	**4.33**	**85.17**	**98.52**	**2.78**

The visual examples shown in Fig. 7(c) further confirm the effectiveness of our DHPF module. The four model variants perform poorly in detecting infrared small targets in low-frequency backgrounds. "$\lambda = [0.1, 0.1, 0.1, 0.1]$" filters out less background information, causing small targets with low signal-to-noise ratio to be interfered with by the complex background. These ultimately lead to both missed targets and false alarms. As the energy filtering ratio increases, more background information is filtered out, and the model detects small targets more accurately, but may also introduce varying degrees of false alarms [shown in the fifth column of the first and second rows and the sixth column of the second row in Fig. 7(c)]. When all energy filtering ratios reach 0.8, excessive filtering results in the removal of small-target information, leading to missed targets [shown in the seventh column in Fig. 7(c)]. This further demonstrates the advantage of our DHPF module in progressively filtering low-frequency information and enhancing small-target representation.

4) Analysis of the Parameter d (Mask Radius) Setting Within the DHPF Module: The parameter d is the mask radius that is adaptively determined based on the aforementioned (7). It can be concluded that the parameter d is not a fixed value and is jointly determined by the current input image frequency energy EC and the preset energy removed ratio λ. In other words, the mask radius d in (7) can satisfy a preset λ, which is used to filter out a specific low-frequency component of the λ ratio of the current input image. At different hierarchical DHPF stages, the mask radius parameter d is determined on the basis of the current input image frequency energy and the current preset λ. As a result, the mask radius parameter d is adaptively varied based on different input images and different preset λ within different DHPF stages.

Furthermore, we also provided a visual example to better illustrate the parameter d setting when calculating E_{removed} in the DHPF module. As shown in Fig. 8, the first row displays the initial input image (IRI) for the 1st DHPF stage, the intermediate input images ($F_{\lambda_1=0.8}$, $F_{\lambda_2=0.4}$, and $F_{\lambda_3=0.2}$) for the 2nd, 3rd, and 4th DHPF stages, respectively. The second row shows the corresponding frequency images within 4 different DHPF stages. The bottom parameter tuples refer to the corresponding preset energy removed ratio λ and the generated mask radius parameter d. From the visual example,

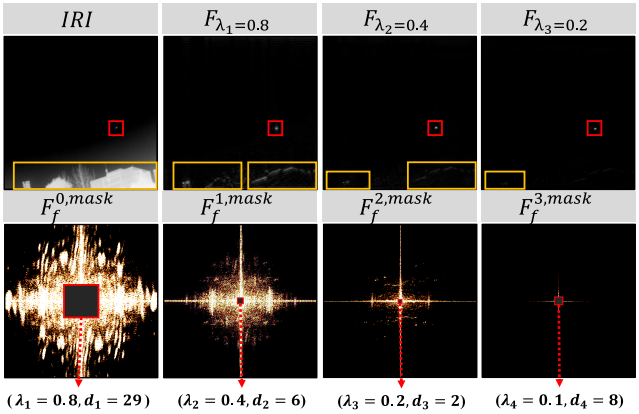

Fig. 8. Visual example to illustrate the setting of the mask radius parameter d within different DHPF stages.

TABLE V

QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS INSIDE THE SLS LOSS. THE BEST RESULTS ARE MARKED IN BOLD

No.	Model Variants	IRSTD-1k			NUDT-SIRST		
		$IoU \uparrow$	$P_d \uparrow$	$F_a \downarrow$	$IoU \uparrow$	$P_d \uparrow$	$F_a \downarrow$
1	$\mathcal{L}_{IoU}$ (wo w)	66.92	90.24	13.36	81.09	97.78	16.36
2	$\mathcal{L}_S$	68.28	94.22	10.80	84.01	98.41	8.48
3	$\mathcal{L}_{IoU}$ + $\mathcal{L}_L$	68.21	91.84	10.75	84.32	98.10	5.29
4	Ours ($\mathcal{L}_{SLS}$)	**70.26**	**94.56**	**4.33**	**85.17**	**98.52**	**2.78**

we can observe that given the preset $\lambda = [\lambda_1, \lambda_2, \lambda_3, \lambda_4] = [0.8, 0.4, 0.2, 0.1]$, the mask radius obtained for different DHPF stages is $d = [d_1, d_2, d_3, d_4] = [29, 6, 2, 8]$. The values of the parameter d generally tend to decrease with decreasing preset λ. However, the calculated values of d do not necessarily follow a decreasing trend with λ decreasing. For example, $F_{\lambda_3=0.2}$ (refers to the input image of the 4th DHPF stage) may contain substantial reduced low-frequency energy, thus a larger value of the mask radius d is needed to satisfy the preset $\lambda_4 = 0.1$. This can explain why d_4 does not decrease but increases to 8. It further demonstrates the dynamic and adaptive characteristics of the DHPF module in filtering low-frequency information.

5) Ablation Experiments Inside $\mathcal{L}_{SLS}$: To validate the effectiveness of $\mathcal{L}_{\text{SLS}}$ in supervising our HDNet, we provided three additional loss variants: 1) "$\mathcal{L}_{\text{IoU}}$" (the loss function is $\mathcal{L}_S$ without w, note that $\mathcal{L}_S$ without w is equivalent to $\mathcal{L}_{\text{IoU}}$); 2) "$\mathcal{L}_S$" (adding w to $\mathcal{L}_{\text{IoU}}$); 3) "$\mathcal{L}_{\text{IoU}}$ + $\mathcal{L}_L$" ($\mathcal{L}_{\text{IoU}}$ with the addition of $\mathcal{L}_L$). The quantitative results are reported in Table V, where the performance metrics of all variants decrease compared with $\mathcal{L}_{\text{SLS}}$. On the IRSTD-1k dataset, we observed that "$\mathcal{L}_{\text{IoU}}$" (Row No. 1) achieves 0.6692 in terms of IoU and 0.9024 in terms of P_d. The "$\mathcal{L}_S$" (Row No. 2) improves these metrics by 1.36% and 3.98%, respectively. Furthermore, "$\mathcal{L}_{\text{IoU}}$ + $\mathcal{L}_L$" (Row No. 3) improves these metrics by 1.29% and 1.60%, respectively. Both "$\mathcal{L}_S$" and "$\mathcal{L}_L$" can effectively reduce F_a. When equipped with both '$\mathcal{L}_S$' and '$\mathcal{L}_L$,' the full "$\mathcal{L}_{\text{SLS}}$" ("$\mathcal{L}_S$" + "$\mathcal{L}_L$," Row 4) achieves the best performance, with IoU and P_d improved by 3.34% and 4.32%, respectively. A similar trend can be observed in the NUDT-SIRST dataset. All these quantitative results demonstrate the effectiveness of "$\mathcal{L}_{\text{SLS}}$."

TABLE VI

COMPARISON OF MODEL COMPLEXITY IN TERMS OF PARAMETERS AND FLOPS BETWEEN OUR HDNET AND 4 MOST RECENT ADVANCED MODELS. "-" INDICATES NO DATA AVAILABLE

Method	Year	Params(M)↓	FLOPs(G)↓	FPS(f/s)↑
UIU-Net [32]	2023	50.54	54.43	9.87
RDIAN [47]	2023	0.217	3.72	180.23
MTUNet [48]	2023	12.75	21.94	10.87
RPCANet [49]	2024	0.68	44.57	2.80
MSHNet [6]	2024	4.07	6.11	68.96
GCI-Net [50]	2024	0.71	55.85	-
SCTransNet [5]	2024	11.19	20.24	34.36
PBT [4]	2024	26.29	28.53	12.25
L^2SKNet [51]	2025	0.899	6.89	110.68
MMLNet [52]	2025	3.58	20.41	38.17
HDNet(Ours)		3.84	5.96	78.81

The visual examples shown in Fig. 7(d) illustrate the advantages of the "$\mathcal{L}_{\text{SLS}}$." The three variants exhibit varying degrees of false alarms and missed targets. "$\mathcal{L}_{\text{IoU}}$" leads to missed targets and false alarms (as shown in the fourth column of the first and second rows). "$\mathcal{L}_S$" results in false alarms in detecting small dim targets (as shown in the fifth column of the second row). "$\mathcal{L}_{\text{IoU}}$ + $\mathcal{L}_L$" leads to significant missed targets (as shown in the sixth column of the first and second rows). However, the "$\mathcal{L}_{\text{SLS}}$," we adopted is capable of detecting targets more comprehensively.

D. Computational Efficiency

We employed network parameters (Params), floating-point operations (FLOPs), and frames per second (FPS) as metrics to quantify model complexity. As shown in Table VI, our model is compared with the most advanced representative methods in the past three years, demonstrating comparable computational efficiency. Notably, our approach maintains a relatively low number of parameters and FLOPs while achieving comparable FPS performance compared to other advanced models. Furthermore, our model exhibits substantially superior accuracy performance compared to the majority of methods listed in Table VI. All these analyses verify that our approach possesses appropriate computational complexity with high accuracy performance.

V. CONCLUSION

In this article, we propose a novel HDNet to tackle the challenges that exist in the IRSTD task. The network consists of two main branches: the spatial-domain branch and the frequency-domain branch. In the spatial domain, the MAC module employs multiple parallel ACCs with different kernel sizes for contrast calculation, enhancing the model's perception of targets with small and variable sizes. In the frequency domain, the DHPF module hierarchically computes the energy of low-frequency signals and dynamically removes specific low-frequency information while preserving the high-frequency information of the image. This allows it to filter out slowly changing, rich low-frequency backgrounds, achieving better preservation and highlighting of the small

targets. Extensive experiments on three public datasets showcase our model's effectiveness and superiority over 26 SOTA models.

REFERENCES

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

[2] J. Zhang and D. Tao, "Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 7789–7817, Nov. 2020.

[3] X. Ying et al., "Local motion and contrast priors driven deep network for infrared small target superresolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5480–5495, 2022.

[4] H. Yang et al., "PBT: Progressive background-aware transformer for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5004513.

[5] S. Yuan, H. Qin, X. Yan, N. Akhtar, and A. Mian, "SCTransNet: Spatial-channel cross transformer network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5002615.

[6] Q. Liu, R. Liu, B. Zheng, H. Wang, and Y. FU, "Infrared small target detection with scale and location sensitivity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17490–17499.

[7] S. Chen, L. Ji, J. Zhu, M. Ye, and X. Yao, "SSTNet: Sliced spatio-temporal network with cross-slice ConvLSTM for moving infrared dim-small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5000912.

[8] W. Duan, L. Ji, S. Chen, S. Zhu, and M. Ye, "Triple-domain feature learning with frequency-aware memory enhancement for moving infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5006014.

[9] W. Duan et al., "Semi-supervised multiview prototype learning with motion reconstruction for moving infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5001215.

[10] S. Chen, L. Ji, W. Duan, S. Peng, and M. Ye, "Motion prior knowledge learning with homogeneous language descriptions for moving infrared small target detection," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2025, vol. 39, no. 2, pp. 2186–2194.

[11] 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, Oct. 1999.

[12] 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.

[13] L. Deng, J. Zhang, G. Xu, and H. Zhu, "Infrared small target detection via adaptive M-estimator ring top-hat transformation," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107729.

[14] 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.

[15] 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.

[16] J. Gao, Z. Lin, and W. An, "Infrared small target detection using a temporal variance and spatial patch contrast filter," *IEEE Access*, vol. 7, pp. 32217–32226, 2019.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] T. Zhang, Y. Fu, and C. Li, "Hyperspectral image denoising with realistic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2228–2237.

[23] 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.*, Jun. 2022, pp. 867–876.

[24] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2022.

[25] 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.

[26] Q. Hou, Z. Wang, F. Tan, Y. Zhao, H. Zheng, and W. Zhang, "RISTDnet: Robust infrared small target detection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[27] Y. Huang et al., "FDDBA-NET: Frequency domain decoupling bidirectional interactive attention network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5004416.

[28] Y. Zhu et al., "Toward robust infrared small target detection via frequency and spatial feature fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 2001115.

[29] T. Ma, G. Guo, Z. Li, and Z. Yang, "Infrared small target detection method based on High-Low-Frequency semantic reconstruction," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[30] 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.

[31] 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.

[32] 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.

[33] M. Zhang, R. Zhang, J. Zhang, J. Guo, Y. Li, and X. Gao, "Dim2Clear network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5001714.

[34] 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.

[35] M. Zhang, Y. Wang, J. Guo, Y. Li, X. Gao, and J. Zhang, "IRSAM: Advancing segment anything model for infrared small target detection," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2024, pp. 233–249.

[36] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 980–993.

[37] J. Guibas, M. Mardani, Z. Li, A. Tao, A. Anandkumar, and B. Catanzaro, "Adaptive Fourier neural operators: Efficient token mixers for transformers," 2021, *arXiv:2111.13587*.

[38] Z. Li et al., "Fourier neural operator for parametric partial differential equations," 2020, *arXiv:2010.08895*.

[39] Z. Huang, Z. Zhang, C. Lan, Z.-J. Zha, Y. Lü, and B. Guo, "Adaptive frequency filters as efficient global token mixers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 6026–6036.

[40] B. Miao, M. Bennamoun, Y. Gao, and A. Mian, "Spectrum-guided multi-granularity referring video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 920–930.

[41] Y. Tatsunami and M. Taki, "FFT-based dynamic token mixer for vision," in *Proc. AAAI Conf. Artif. Intell.*, Jan. 2023, pp. 15328–15336.

[42] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[43] 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.

[44] L. Zhang, L. Peng, T. Zhang, S. Cao, and Z. Peng, "Infrared small target detection via non-convex rank approximation minimization joint $l_{2,1}$ norm," *Remote Sens.*, vol. 10, no. 11, p. 1821, Nov. 2018.

[45] M. Zhang et al., "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.

[46] 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.
[47] 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, Art. no. 5000513.
[48] T. Wu et al., "MTU-Net: Multilevel TransUnet for space-based infrared tiny ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601015.
[49] F. Wu, T. Zhang, L. Li, Y. Huang, and Z. Peng, "RPCANet: Deep unfolding RPCA based infrared small target detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2024, pp. 4797–4806.
[50] M. Zhang, K. Yue, B. Li, J. Guo, Y. Li, and X. Gao, "Single-frame infrared small target detection via Gaussian curvature inspired network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5005013.
[51] F. Wu, A. Liu, T. Zhang, L. Zhang, J. Luo, and Z. Peng, "Saliency at the helm: Steering infrared small target detection with learnable kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5000514.
[52] Q. Li, W. Zhang, W. Lu, and Q. Wang, "Multibranch mutual-guiding learning for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5605710.
[53] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst.*, 2017, pp. 5998–6008.

Mingzhu Xu (Member, IEEE) received the B.S., M.Sc., and Ph.D. degrees from Harbin Institute of Technology (HIT), Harbin, China, in 2013, 2015, and 2021, respectively.

He is currently an Assistant Professor with the School of Software, Shandong University, Jinan, China. His research interests include computer vision, multimedia computing, and information retrieval.

Dr. Xu is also an Invited Reviewer for prestigious journals, including IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *Information Science*, ACM MM, NeurIPS, and ICML.

Chenglong Yu is currently pursuing the M.Sc. degree with the School of Software, Shandong University, Jinan, China.

His research interests include computer vision and infrared small-target detection.

Zexuan Li is currently pursuing the B.S. degree with the School of Software, Shandong University, Jinan, China.

His research interests include computer vision and infrared small-target detection.

Haoyu Tang (Member, IEEE) received the B.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2016 and 2021, respectively.

He is currently an Assistant Professor with the School of Software, Shandong University, Jinan, China. His research interests include machine learning and multimedia retrieval.

Yupeng Hu (Member, IEEE) received the Ph.D. degree in software engineering from Shandong University, Jinan, China, in 2018.

He is currently an Associate Professor with the School of Software, Shandong University. Various parts of his work have been published in famous journals and forums, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, *Science China Information Sciences*, and *ACM Multimedia*. His research interests include information retrieval and data mining.

Dr. Hu has served as a PC member for ACM MM, ACL, and AAAI and a reviewer for IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING and IEEE TRANSACTIONS ON MULTIMEDIA.

Liqiang Nie (Senior Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong University, Xi'an, China, in 2009, and the Ph.D. degree from the National University of Singapore (NUS), Singapore, in 2013.

He is currently the Dean with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen Campus), Shenzhen, China. He has co-authored more than 100 CCF-A papers and five books, with 26k plus Google Scholar citations. His research interests include multimedia content analysis and information retrieval.

Dr. Nie is a member of the ICME Steering Committee. He is a fellow of AAAI and IAPR. Meanwhile, he is the regular Area Chair or SPC of ACM MM, NeurIPS, IJCAI, and AAAI. He has received many awards over the past three years, like ACM MM and SIGIR best paper honorable mention in 2019, the AI 2000 most influential scholars 2020, the SIGMM rising star in 2020, the MIT TR35 China 2020, the DAMO Academy Young Fellow in 2020, the SIGIR Best Student Paper in 2021, first price of the provincial science and technology progress award in 2021 (rank 1), and provincial youth science and technology award in 2022. Some of his research outputs have been integrated into the products of Alibaba, Kwai, and other listed companies. He is an Associate Editor of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Information Science*.