# Dynamic Context-Aware Pyramid Network for Infrared Small Target Detection

Xiaolong Chen, Jing Li, Tan Gao, Yongjie Piao, Haolin Ji, Biao Yang, and Wei Xu

***Abstract*—Detecting faint and diminutive infrared targets devoid of clearly defined shape and texture details in intricate surroundings remains a formidable challenge within the domain of target detection. Current methodologies employing deep neural networks and pooling operations can easily cause small target loss, resulting in suboptimal detection outcomes. To address these issues, we design an innovative dynamic context-aware pyramid network. It comprises three core modules: dynamic context modulation (DCM), dynamic pyramid context (DPC), and shuffle attention fusion (SAF). Specifically, the DCM module is designed to adaptively capture diverse-scale information from input images, adapting to various target dimensions, and enhancing the feature representation capabilities crucial for effective target detection. Subsequently, the DPC module adaptively captures multiscale features and better utilizes contextual information by aggregating multiple DCM modules. This facilitates the retention of essential semantic information about small infrared targets within deeper network layers. Finally, through the designed SAF module, we facilitate the exchange of information within the same layer and establish correlations between different layers, ensuring the fusion of shallow spatial positional information and deep semantic information to enhance the overall detection performance. Furthermore, comprehensive ablation studies are conducted to substantiate the efficacy of the designed modules within the proposed network architecture. Simultaneously, we conducted a comparative analysis of the proposed network algorithm against several state-of-the-art methodologies for infrared small target detection, employing multiple evaluation metrics. The results consistently demonstrated the proposed model attains superior detection performance on the publicly available IRSTD-1 k, SIRST-Aug, and NUDT-SIRST datasets.**

***Index Terms*—Context-aware pyramid, deep learning, dynamic convolutional, infrared small targets, shuffle attention (SA).**

## I. Introduction

INFRARED imaging is capable of penetrating rain and fog interference, providing clear images of targets in low-light conditions, and maintaining excellent concealment [1]. Consequently, as a significant research focus within computer vision, infrared small target detection technology has found widespread practical utilities across various domains, such as early warning systems, medical imaging, maritime rescue, and space target monitoring [2], [3]. Nevertheless, in contradistinction to visible light imaging, the infrared image is beset by intrinsic limitations, most notably lower contrast and resolution. The considerable distance between infrared sensors and the targeted objects results in these objects typically occupying only a few pixels within the image, thereby exacerbating the challenges associated with extracting features. Moreover, infrared targets lack discernible shape and texture, while being compounded by a lower signal-to-noise ratio, often causing them to be submerged in intricate backgrounds and sensor-generated noise, as depicted in Fig. 1. Consequently, the development of robust and efficient methodologies for detecting diminutive targets represents a highly significant research topic within the realm of target detection.

Presently, infrared diminutive target detection algorithms can be broadly categorized into two major classes: multiframe and single-frame methods. The multiframe method leverages both temporal and spatial information from multiple images to detect targets, but often exhibits high complexity and low execution efficiency, rendering it impractical for real-time end-to-end detection [4]. In contrast, single-frame detection methodologies employ techniques such as target enhancement or background suppression to extract features, such as grayscale, contrast, and gradients from infrared images, thereby enabling target detection. Traditional single-frame detection strategies encompass three primary paradigms: background suppression to heighten the discernibility of small targets [5], [6]; augmentation of contrast between the background and target neighborhood regions to achieve direct detection [7], [8], [9], [10]; and the segregation of the linear superposition of target and background images by formulating the detection problem as a matrix optimization and decomposition problem [11], [12], [13], [14]. However, these methods often rely on prior models, necessitating manual parameter adjustment, exhibit limitations in recognizing nuanced scene changes, and manifest sensitivity to hyperparameters. Consequently, when real-world scenarios change, these methods yield detection outcomes that lack stability and robustness.

In recent years, many excellent deep learning-based methods have emerged in the field of remote sensing image interpretation, such as large kernel sparse convnet weighted by multifrequency attention [15], multistage information complementary fusion network [16], and spatial–spectral perception network [17].

Manuscript received 13 May 2024; revised 4 July 2024; accepted 21 July 2024. Date of publication 29 July 2024; date of current version 9 August 2024. *(Corresponding author: Wei Xu.)*

Xiaolong Chen, Tan Gao, Yongjie Piao, Haolin Ji, Biao Yang, and Wei Xu are with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, Changchun 130033, China, also with the University of Chinese Academy of Science, Beijing 100039, China, also with the Key Laboratory of Space-Based Dynamic and Rapid Optical Imaging Technology, Chinese Academy of Sciences, Changchun 130033, China, and also with the Jilin Provincial Key Laboratory of Aerospace Advanced Optical Imaging Technology, Jilin 130033, China (e-mail: chenxiaolong20@mails.ucas.ac.cn; gaotan19@mails.ucas.ac.cn; pyj0314@163.com; jihaolin21@mails.ucas.ac.cn; yangbiao211@mails.ucas.ac.cn; xwciomp@126.com).

Jing Li is with the Beijing Institute of Tracking and Communication Technology, Beijing 100094, China (e-mail: zuobin97117@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3434330

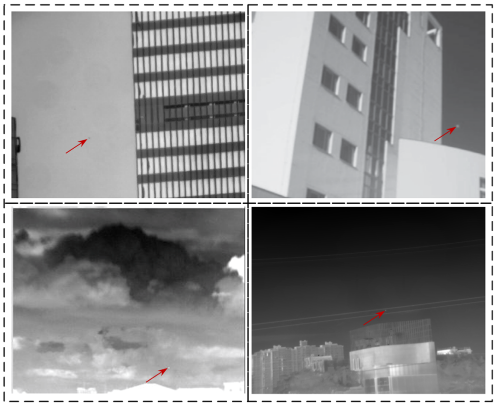


Fig. 1. Example of submerged infrared small targets.

Meanwhile, the proliferation of publicly available datasets [18], [19], [20] and the evolution of deep learning within the realm of computer vision have propelled a heightened focus on the application of convolutional neural networks (CNNs) for diminutive target detection. Numerous inventive detection algorithms grounded in the principles of deep learning have emerged in response to this pursuit. Liu et al. [21] were pioneers in introducing a CNN-based algorithm for the detection of infrared faint and diminutive targets. Notably, in intricate scenarios where small targets coexist with elements, such as buildings, interfering objects, and sky, the performance of deep learning-based detection algorithms is typically superior to that of conventional detection methods. By improving the signal-to-noise ratio in segmentation results and extracting spatial and frequency contexts, Zhang et al. [22] proposed a novel Dim2Clear network (Dim2Clean) for infrared small target detection. Due to the low contrast between small targets and noisy backgrounds in infrared images, Zhang et al. [23] also proposed a model with feature compensation and cross-level correlation, a thermodynamics-inspired multi-feature network [24], and a novel RKformer model with an encoder-decoder structure [25]. Dai et al. [26] contributed an asymmetric context module (ACM) designed to accentuate the semantic information of small targets through the utilization of multilevel features. The attention-guided pyramidal context network (AGPCNet) [19] applied attention mechanisms for feature enhancement during deep feature processing. Aiming at the low computational efficiency of existing models, Zhang et al. [27] designed a novel wavelet regularized soft channel pruning method to establish an efficient IRPruneDet model. Nevertheless, the model performance is encumbered by an abundance of hyperparameters necessitating experimental tuning, resulting in weak generalization across diverse datasets. Wang et al. [28] introduced an internally attention-aware network designed characterized by a coarse-to-fine strategy to suppress false alarm sources. Zhang et al. [20] devised edge blocks and bidirectional attention aggregation blocks inspired by Taylor finite differences, providing a mathematical underpinning for the precise delineation of target shapes. Li et al. [29] developed a dense nested attention network (DNANet) that comprehensively integrates context information for small targets through multilevel feature enhancement and fusion. To address the issue of imbalance between backgrounds and small targets, Yang et al. [30] designed an adaptive threshold focal loss function, which automatically adjusts the loss weights to enhance the learning capability of small target features. Wu et al. [31] proposed an interpretable infrared dark target detection deep network (RPCANet) by converting the small target detection task into sparse target extraction, low-rank background estimation, and image reconstruction. Despite the strides made by these CNN-based methodologies in enhancing small target detection, extant networks grapple with challenges including inadequacies in feature representation, target loss, and low detection accuracy.

In conclusion, the detection of faint and diminutive targets is challenged by factors, such as limited pixel occupancy, indeterminate shapes, unclear contours, complex background variations, and low resolution. To address these challenges, an effective detection model not only recurrently exploits intralayer channel and spatial information but also incorporates cross-layer feature fusion strategies to establish mutual connections between feature representations from different layers. Drawing inspiration from prior research [32], [33], we propose a novel solution named the dynamic context-aware pyramid network (DCAPNet) to address the aforementioned issues. The network designed to enhance detection efficacy without increasing computational complexity, integrates a DPC module and an SAF module. Through these modules, DCAPNet establishes inter-element correlations, resolves conflicting information interference, and augments semantic information of small targets, thereby improving overall detection performance.

This article makes several significant contributions as follows.

1) We replace the traditional pyramid pooling operation with the DCM module, which not only prevents the loss of detailed target information but also dynamically captures multiscale information. This adaptation makes the proposed DCAPNet algorithm more suitable for situations with complex background variations.
2) An innovative dynamic pyramid context (DPC) module is introduced in this study, leveraging a dynamic context modulation (DCM) module for the adaptive enhancement of contextual information across multiple hierarchical features. The fusion of original features with representations from multiple DCM modules is employed to acquire information on distinct scales in infrared images, thereby enriching the substantive content and semantic information of feature representations.
3) A meticulously designed shuffle attention fusion (SAF) module is presented to effectively suppress redundant and conflicting information, while concurrently fusing detailed and semantic information. This module not only focuses on intralayer channel and spatial information but also facilitates feature fusion in both the bottom-up and top-down directions. The resulting comprehensive feature integration significantly enhances the detection performance of the proposed model.

The rest of this article is organized as follows. Section II provides a concise overview of related work. Section III elaborates on the network structure of the proposed DCAPNet. In

Section IV, we present the relevant experimental results and make quantitative and qualitative analyses. Section V is the discussion section. Finally, Section VI concludes this article.

## II. Related Works

### A. Pyramid Structure

In addressing the issue of insufficient scale variation, the construction of a multiscale pyramid structure is a prevalent strategy within the domain of object detection. The CNN-based feature pyramid network [34] is a top-down network structure with lateral connections, capable of producing multiscale feature representations. And the feature representation of each layer is predicted individually. Nonetheless, the intrinsic semantic disparities among features of disparate scales necessitate careful consideration. The direct fusion of features across scales introduces copious redundancies and conflicting information, thereby diminishing the expressive capacity of multiscale representations. The atrous spatial pyramid pooling (ASPP) module, initially propounded in DeepLabV2 [35], consists of four parallel dilated convolution modules with varying rates, aiming to acquire a dense receptive field and extract multiscale contextual information. In DeepLabV3 [36], improvements were made to the ASPP module by adjusting the dilation rates and introducing global pooling operations for the purpose of capturing global features. Subsequently, DeepLabV3+ [37] opted for the substitution of standard convolutions in ASPP with depthwise convolutions to reduce parameter volume and enhance computational efficiency. However, the utilization of such dilated convolutions has been associated with information loss of small targets, engendering grid artifacts and boundary effects. Furthermore, the pooling pyramid network module proposed in PSPNet [38] comprises four parallel adaptive pooling channels. This module endeavors to capture multiscale contextual information through divergent pooling operations but may incur some loss of fine-grained details.

### B. Attention Mechanism

Given the limited availability of discriminative features for small targets, the task of detecting diminutive targets remains a formidable challenge within the domain of computer vision [39]. Therefore, introducing the attention mechanism enables CNN to concentrate more on important information in the images, thereby improving the detection performance and generalization capability of the network. The self-attention mechanism [40] emerges as an adaptive strategy for discerning interdependencies between elements by calculating their relative importance. An intricate variant of this, the multihead attention mechanism [41], employs multiple self-attention matrices and associated weight matrices for weighted aggregation and concatenation of results, thereby augmenting the expressive power of the mode. SENet [42] evaluates the importance levels of each channel to elevate the capability of the network in feature representation. Similarly, the CBAM [43] module integrates spatial attention and channel attention and deftly directs focus toward pivotal image regions while disregarding less consequential areas. Consequently, how to apply attention mechanisms to accentuate regions housing small infrared targets remains an exigent concern demanding meticulous investigation and resolution.

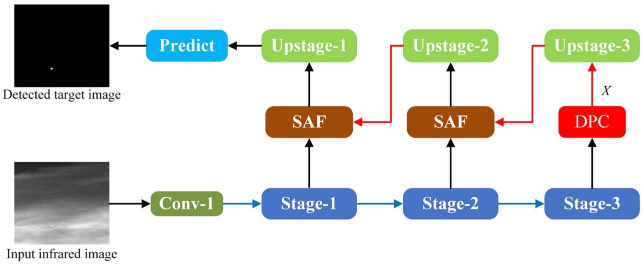


Fig. 2. Overall architecture of DCAPNet, which incorporates a DPC module and two SAF modules into a U-Net (the red and blue lines represent upsampling and downsampling operations, respectively, while stage and upstage represent the feature layers after downsampling and upsampling, respectively, red and blue lines represent upsampling and downsampling operations, respectively).

### C. Infrared Small Target Detection

Various researchers have proposed numerous detection networks tailored for the identification of small objects within the visible light spectrum in recent years [44], [45]. However, the direct application of these networks to infrared small target detection tasks may lead to inaccurate results, primarily arising from the bounding box detection methods used in visible light images that significantly differ from those suitable for infrared small targets. Nevertheless, the recent proliferation of publicly available infrared small target datasets has catalyzed a paradigm shift in infrared small target detection techniques, transitioning from traditional handcrafted feature extraction algorithms to more advanced machine learning and deep learning methodologies. Concurrently, the CNN-based pixel-level segmentation methods [46], [47] have been employed for infrared small target detection tasks, yielding notable detection outcomes in various instances, such as ACM [26], TBC-Ne [48], and DNANet [29]. However, the issue of information loss concerning small targets persists in deep networks, resulting in diminished robustness when confronted with complex background variations and the dynamic, faint nature of small targets. The intricacies surrounding information loss in the context of small targets within deep models underscore the ongoing imperative to refine and advance detection methodologies for infrared small targets.

## III. Methodology

In this section, we first provide a detailed exposition of the overall framework of the proposed DCAPNet, as illustrated in Fig. 2. Subsequently, we introduce the components of the dynamic contextual modulation module. Following that, we elaborate on how the DPC module integrates the multiscale features generated by the DCM module. Finally, we employ the SAF module to effectuate a bidirectional fusion of low-level and high-level features. The section aims to provide a nuanced understanding of the structural intricacies of the proposed model and its interplay of components.

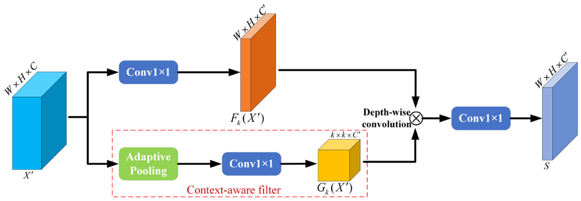


Fig. 3. DCM module.

### A. Network Architecture

Due to the scarcity of pixels and the diminutive size of infrared targets, a multiscale representation proves to be an efficacious method for extracting features pertinent to small infrared targets. While prior methodologies, such as the utilization of dilated convolutions to expand the receptive field [19], have been employed to construct multiscale representations, these convolutional operations bring forth a substantial computational burden and are prone to inducing information loss within local neighborhoods. The two-stage network needs to generate a large number of region proposals, and the network layer number is large and the calculation amount is large, which makes it easy to lose small targets in the process of feature extraction. The one-stage network does not have the step of generating region proposals, and the running speed is fast, which can meet the real-time requirement of detection. Meanwhile, in order to prevent small targets from being lost in deep networks and realize efficient end-to-end detection, we select a one-stage detector. To confront these challenges, this study proposes a DCAPNet, the overarching architecture of which is illustrated in Fig. 2.

Considering that small targets occupy a very small proportion in the infrared image and most of the images are redundant background regions, we adopt the DCM module to replace the traditional pooling operation in the pyramid, which can prevent the loss of target information. Meanwhile, since the target lacks texture information and is extremely weak, we propose a DPC module that aggregates multiple DCM modules. As shown in Fig. 3, the context-aware filter embeds rich content and high-level semantic information, and can adapt to the input image to capture different size information inside the image. Each DCM module can handle proportional changes related to input size. Therefore, the DPC module that aggregates multiple DCM modules is a dynamic convolution module. This module adaptively learns suitable convolution kernels to better utilize contextual information and capture multiscale semantic information, thereby adaptively learning information of different sizes in the image. Aiming at the problem that small targets are easily submerged by complex changing backgrounds and easily lost in deep networks, the idea of information exchange at the same level and the combination of high- and low-level features is put forward. For this purpose, we design an SAF module that aggregates local and global information and fully fuses multilayer features to retain and highlight small infrared targets in the high-level features.

DCAPNet is composed of a backbone convolutional network, an innovative DPC module, and two innovative SAF modules. The DPC module consists of multiple existing DCM modules. The SAF module consists of a shuffle attention (SA) module and an original bidirectional asymmetric fusion (BAF) module. Upon receiving an infrared image as input, DCAPNet initiates a feature extraction process using a ResNets network as the backbone, yielding the feature map denoted as $X \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}}$. Subsequently, the DPC module adaptively subdivides semantic information to generate a multiscale contextual understanding. Finally, the SAF module facilitates the effective fusion of semantically enriched deep-layer features with spatially refined shallow-layer features, culminating in the acquisition of more precise spatial and positional information pertinent to small infrared targets.

### B. Dynamic Contextual Modulation Module

As shown in Fig. 3, the DCM modules adaptively acquire specific ratio representations correlated with input feature images. The red dashed box in the figure is referred to as the context-aware filter. The filter is adaptively generated from the contextual information of the input feature map $X' \in \mathbb{R}^{W\times H\times C}$, where $W$, $H$, and $C$ denote the width, height, and channel number of the feature map, respectively. In comparison to traditional filters, these adaptive filters encompass rich contextual information and elevated semantic acumen. They are capable of better adapting to the input feature map, capturing diverse size information of the input image, thereby exhibiting heightened adaptability and flexibility. Therefore, we apply them to representations at different feature scales for multiscale learning. In the following, we present a detailed exposition of the DCM module.

Initially, the feature map $X' \in \mathbb{R}^{W\times H\times C}$ extracted from the backbone network is bifurcated into two distinct branches for subsequent processing. In branch one, the map $X'$ undergoes a $1 \times 1$ convolution aimed at channel dimensionality reduction, thereby simplifying the feature map to $F_k(X') \in \mathbb{R}^{W\times H\times C'}$, where $C'$ ($C' \leq C$) denotes the channel count in the simplified feature map, and $k$ signifies the convolution kernel size. Branch two sequentially executes adaptive average pooling operations and $1 \times 1$ convolution operations on the input feature map, generating a filter $G_k(X') \in \mathbb{R}^{k\times k\times C'}$ with a kernel size of $k$. The subsequent fusion of feature maps from both branches is achieved through a depthwise convolution operation, yielding a feature map tailored to a specific scale

$$S_k = F_k(X') \otimes G_k(X') \tag{1}$$

where the symbol $\otimes$ denotes the depthwise convolution operation. Finally, the $1 \times 1$ convolution operation is performed on $S_k$ to fuse the channel information of the upper and lower branch feature maps, so as to obtain the specific scale feature map $S \in \mathbb{R}^{W\times H\times C'}$ output by the DCM module.

Based on the analysis provided previously and the information depicted in Fig. 3, it is evident that in a DCM module with a specific filter kernel size $k$, the generation process of a $k \times k$ context-aware filter unfolds as follows: initially, an adaptive average pooling operation is executed on the map $X'$ to derive a $k \times k$ region based on contextual information, succeeded by a subsequent $1 \times 1$ convolutional processing step. Remarkably, this filter necessitates only a single $1 \times 1$ convolution operation.

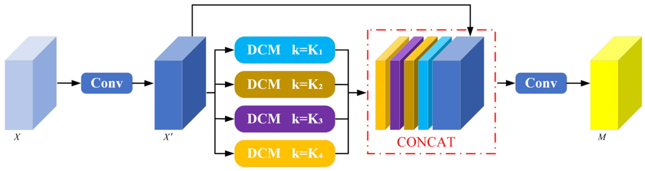


Fig. 4. DPC module.

Consequently, we can employ multiple DCM modules with context-aware filters featuring distinct kernel sizes to replace the traditional pooling operation, generating multiscale feature representations. This module utilizes dynamic multiscale filters for semantic segmentation, preventing the loss of target detail information while dynamically capturing multiscale details.

### *C. DPC Module*

The pyramid pooling module is adept at constructing global scene prior information on the ultimate layer feature map within deep neural networks, thereby mitigating the loss of contextual information across diverse subregions [38]. However, pooling operations inherently lead to the loss of nuanced details. Therefore, we construct a DPC module after the deepest feature layer, which can dynamically obtain multiscale context information by using context-aware filters with different convolution kernel sizes.

The prime objective of the DPC module lies in the aggregation of multiscale feature representations emanating from multiple DCM modules, as illustrated in Fig. 4. First, we perform feature extraction on the image $X$ via convolutional operations in the backbone network, yielding the feature map $X'$. Subsequently, we concurrently execute DCM modules with multiple filters featuring distinct kernel sizes $k$. Following $1 \times 1$ convolutional processing, the dimension-reduced results $S = [S^{K_1}, S^{K_2}, S^{K_3}, \ldots]$ are obtained, where $K$ represents the kernel size of the context-aware filter. The feature maps $[S^i]$ generated by multiple DCM modules are then concatenated with the map $X'$ extracted from the backbone network. Finally, channel information from the aggregated feature is processed through a $1 \times 1$ convolutional operation, yielding the output result $M$ of the DPC module. Thus, DCM modules of different scales form a contextual pyramid module.

### *D. SAF Module*

Deep neural networks excel at extracting advanced semantic information for small targets. Nevertheless, diminutive targets occupy only a limited number of pixels and lack intrinsic information, and targets may be susceptible to loss in deep layers of the network. Therefore, how to prevent the reduction or even loss of the fine detail information of the target in the deep network becomes a critical concern. Simultaneously, features of different scales exhibit significant semantic differences. The direct fusion of different scale features will introduce a surfeit of redundant and conflicting information, which will diminish the capability of multiscale expression. As shown in Fig. 6, to address these challenges, we propose an SAF module that consists of an SA module and a BAF module. The SA module effectively integrates channel and spatial attention mechanisms, preventing minute target features from being submerged in conflicting information while emphasizing the most salient regions of targets. Meanwhile, the BAF module consists of a bottom-up local attention modulation branch and a top-down global attention modulation branch, aiming to fuse spatial detail and semantic information. In addition, as illustrated in Fig. 2, the proposed SAF module bidirectionally integrates information from each downsampled feature layer, preserving both target semantic context and detailed information, effectively addressing the issue of small target loss. Subsequently, we provide detailed descriptions of the SA module and the BAF module.

SA module: the attention mechanism enables neural networks to focus more on finding relevant and significant information in the input data for the current output while suppressing irrelevant information, thereby enhancing the performance and generalization capability of the network. Zhang et al. [32] proposed an efficient Shuffle unit, which is an ultralightweight attention mechanism effectively integrating spatial and channel attention types. As depicted in Fig. 5, this unit not only augments the efficacy of network feature representation but also accelerates computational speed by mitigating model complexity and reducing network parameters. We combine the SA module with the BAF module to fully exploit the spatial and channel dependencies of features across low and high-level networks to enhance the representation ability of target features and suppress conflict information and noise.

Assuming the input feature map is denoted as $x \in \mathbb{R}^{w\times h\times c}$, where $c$, $w$, and $h$, respectively, represent the channel number, height, and width of the map $x$, SA partitions the input map into $g$ groups along the channel dimension: $x = [x_1, x_2, \cdot\cdot\cdot, x_g]$, $x_k \in \mathbb{R}^{w\times h\times \frac{c}{g}}$. For each group of subfeatures $x_k$, we need to generate distinct importance coefficients through the spatial and channel attention modules, so that they gradually learn specific semantic features. The subgroup $x_k$ is subsequently bifurcated along the channel dimension into two branches: $x_{k1}, x_{k2} \in \mathbb{R}^{w\times h\times \frac{c}{2\tilde{}g}}$. The upper branch is dedicated to channel attention feature learning: to be as lightweight as possible, we employ global average pooling (GAP) to first capture global information within the subfeature $x_{k1}$, yielding the channelwise statistical map denoted as $l \in \mathbb{R}^{1\times 1\times \frac{c}{2\tilde{}g}}$. Subsequently, a series of operations involving a scaling transformation function $F_c \in \mathbb{R}^{1\times 1\times \frac{c}{2\tilde{}g}}$ and a sigmoid activation function are executed in turn. Finally, an elementwise multiplication is performed between each subgroup feature $x_{k1}$ and the computed attention map to generate the result $x'_{k1} \in \mathbb{R}^{w\times h\times \frac{c}{2\tilde{}g}}$ representing the channel branch. The computational process for generating channel attention features is delineated as follows:

$$l = F_{gp}(x_{k1}) = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} x_{k1}(i, j) \tag{2}$$

$$x'_{k1} = \sigma\left(F_c(l)\right) \cdot x_{k1} = \sigma\left(W_1 l + b_1\right) \cdot x_{k1} \tag{3}$$

where $\sigma(\cdot)$ represents the sigmoid function, $W_1 \in \mathbb{R}^{1\times 1\times \frac{c}{2\tilde{}g}}$ and $b_1 \in \mathbb{R}^{1\times 1\times \frac{c}{2\tilde{}g}}$ are the parameters of $l$ scaling and moving,

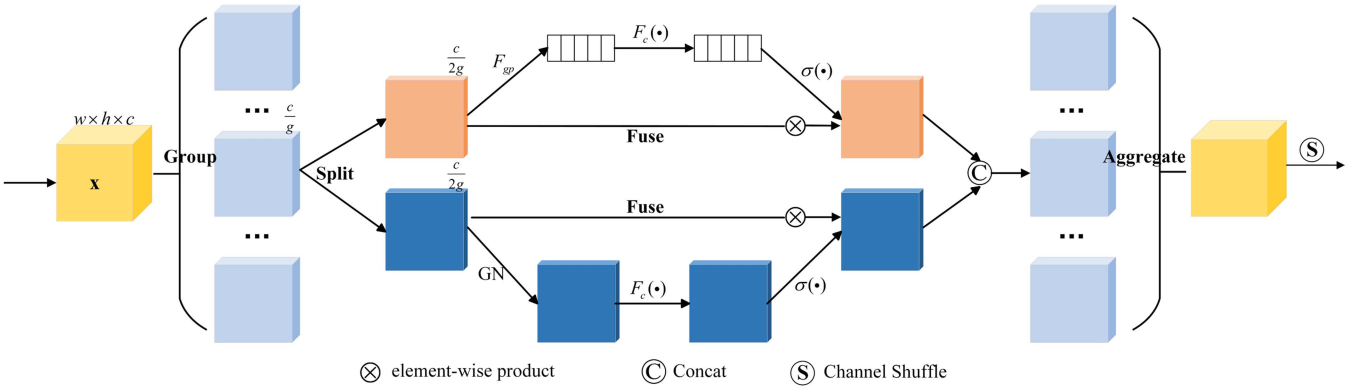


Fig. 5. SA module ($\sigma(\cdot) = \text{sigmoid}(\cdot)$, $F_c(x) = Wx + b$).

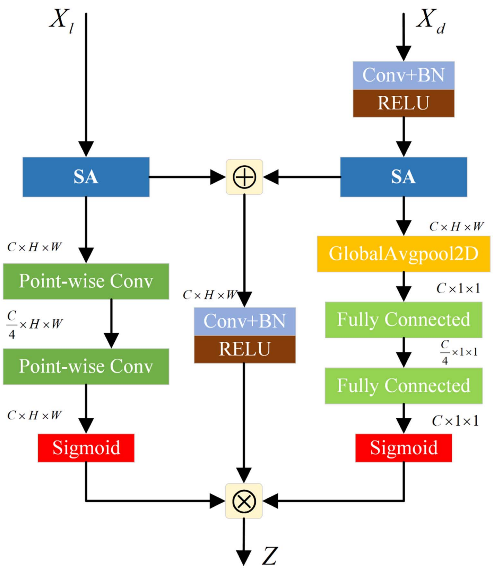


Fig. 6. SAF module.

respectively. The lower branch is dedicated to learning spatial attention features, which is more concerned with the crucial factor of "where," complementing the emphasis of channel attention on "what." In the implementation, the group norm (GN) [49] is first applied to process the input subfeature $x_{k2}$. Subsequently, feature representation of $l$ is enhanced through a scaling transformation function $F_c \in \mathbb{R}^{1\times 1\times \frac{c}{2^* g}}$. Finally, an elementwise multiplication is performed with the bifurcated sub-feature $x_{k2}$, resulting in the output of the spatial branch denoted as $x'_{k2} \in \mathbb{R}^{w\times h\times \frac{c}{2^* g}}$. This process can be described as follows:

$$\begin{aligned} x'_{k2} &= \sigma\left(F_c\left(\mathrm{GN}(x_{k2})\right)\right) \cdot x_{k2} \\ &= \sigma\left(W_2 \mathrm{GN}(x_{k2}) + b_2\right) \cdot x_{k2} \end{aligned} \quad (4)$$

where $W_2$ and $b_2$ are parameters of shape $\mathbb{R}^{1\times 1\times \frac{c}{2^* g}}$. Following this, the outputs from the upper and lower branches are intricately concatenated to generate a feature map $x'_k$ with the equivalent number of channels as the subfeature $x_k$, as expressed by the following equation:

$$x'_k = [x'_{k1}, x'_{k2}] \in \mathbb{R}^{w\times h\times \frac{c}{g}}. \quad (5)$$

Next, all the subfeatures $x'_k$ obtained after parallel processing are aggregated to produce a feature map of the same scale as the feature $x$. Ultimately, information exchange between different groups on the channel is realized through the "channel shuffle" operator in ShuffleNet v2 [50].

BAF module: shallow features retain spatial details of the target while deep features have rich semantic information. Therefore, ensuring the extraction of deep semantics for targets while retaining fine-grained details is a pivotal challenge. ACM proved that the BAF module performs better than the top-down feature fusion module. The module comprises a bottom-up mechanism that handles low-level detail information using a pointwise local channel attention mechanism and a top-down mechanism that addresses high-level semantic information using a global channel attention mechanism. Assuming the low-level features are denoted as $X_l$ and the high-level features as $X_d$, with $X_d$ being resized to match the dimensions of $X_l$ through convolution. The computation formulations for the global channel and spatial attention mechanisms of this module are expressed by (6) and (7), respectively,

$$L\left(X_d\right) = \sigma(\Im(W_2\delta(\Im(W_1 x_d)))) \quad (6)$$

$$G\left(X_l\right) = \sigma(\Im(\mathrm{PWConv}_2(\delta(\Im(\mathrm{PWConv}_1(X_l)))))) \quad (7)$$

where $\delta$, $\sigma$, and $\Im$ represent the rectified linear unit, the sigmoid function, and batch normalization, respectively. PWConv denotes pointwise convolution [51]. $x_d$ is the channel statistics obtained by GAP: $x_d = \frac{1}{H\times W}\sum_{i=1,j=1}^{H,W} X_d[:, i, j]$. However, the module inadequately utilizes intralayer feature information, making the detailed information of the target susceptible to being overwhelmed by background noise.

To solve problems, such as the loss of fine details in small targets, interlayer information redundancy, and insufficient feature fusion, we incorporate the SA module into the BAF module to construct the SAF module. The SA attention modulation of the details and semantic information of the target is performed by the SAF module. Modulation facilitates the flow of spatial and channel information within the same layer, ensuring that fine details of the target are preserved without being obscured by noise. Concurrently, the SAF module employs a local point attention mechanism and a global channel attention mechanism to perform a cross-layer fusion of distinct features, which ensures the integration of deep semantics and shallow spatial positional information.

Given the high-level feature $X_d$ and the low-level feature $X_l$ hey share the same scale after upsampling, as illustrated in Fig. 6. First, the SA module independently processes the input high-level and low-level features to obtain ${X_d}'$ and ${X_l}'$. Then, the top-down global channel modulation module and the bottom-up local point attention module are executed in parallel to obtain the feature maps ${X_d}''$ and ${X_l}''$. The calculation process is defined in (8) and (9), respectively

$$ {X_d}'' = L({X_d}') = L(\mathrm{SA}(X_d)) \tag{8} $$

$$ {X_l}'' = G({X_l}') = G(\mathrm{SA}(X_l)) \tag{9} $$

where $L(\cdot)$, $\mathrm{SA}(\cdot)$, and $G(\cdot)$ denote global channel modulation, SA modulation, and local point attention modulation, respectively. Subsequently, the ${X_d}'$ and ${X_l}'$ are added and restored to the scale of $W \times H \times C$, and then the product with the processed feature maps ${X_d}''$ and ${X_l}''$ pixel by pixel to gain the final output feature $Z \in \mathbb{R}^{C\times H\times W}$

$$ \begin{aligned} Z &= \delta\left(W({X_l}' + {X_d}')\right) \otimes {X_l}'' \otimes {X_d}'' \\ &= \delta\left(W(\mathrm{SA}(X_l) + \mathrm{SA}(X_d))\right) \otimes L(\mathrm{SA}(X_l)) \otimes G(\mathrm{SA}(X_d)) \end{aligned} \tag{10} $$

where $\otimes$ represents elementwise multiplication.

## IV. Experiments and Analysis

We systematically assess the proposed DCAPNet through both qualitative and quantitative analyses in this section. We commence by elucidating the details of the network implementation. Following that, we elaborate on the experimental setup, encompassing baseline methods, datasets, and evaluation metrics. Subsequently, to assess the effectiveness of each constituent module within DCAPNet, an exhaustive ablation study is carried out. Finally, we subject our DCAPNet to both subjective and objective comparisons with various classical detection algorithms.

### A. Implementation Details

We converted the diminutive target detection task into a semantic segmentation problem, utilizing the U-Net network architecture with ResNets [52] as the segmentation backbone. The devised network architecture incorporates the SoftIoU [53] function as the designated loss function, and optimization is achieved via stochastic gradient descent with weight decay coefficients and momentum set to 0.9 and 0.0004, respectively. The experimental setup specifies a batch size of 8, an initial learning rate of 0.05, and a total of 50 training iterations for the network. Furthermore, we implement a poly decay strategy for learning rate reduction, wherein the learning rate is multiplied by $(1 - \frac{\text{iter}}{\text{total\_iter}})^{0.9}$ after each iteration. The computational framework for all network algorithms is instantiated on a computational system featuring an Intel Core i9-10900 CPU @ 2.80 GHz and Nvidia GeForce RTX 3070 GPU. And the implementation is conducted utilizing the PyTorch framework.

### B. Experimental Settings

Datasets: Given the restricted scale of the publicly accessible SIRST dataset [26], consisting of a mere 427 infrared images, we leverage the augmented SIRST dataset, denoted as SIRST-Aug, as introduced by Zhang et al. [19]. The augmented dataset comprises 545 testing images and 8525 training images. Furthermore, we evaluate the detection capabilities of the proposed DCAPNet using the IRSTD-1 k dataset [20] and the NUDT-SIRST dataset [29], which feature diverse target shapes, sizes, and scenes. The image sizes of the three datasets are different, which cannot match the infrared data of the actual scene, making it difficult for the network to uniformly process. To address the variance in image sizes between the two datasets, we fix the input size of DCAPNet images to $256 \times 256$, aligning them more closely with the dimensions of real-world infrared images.

Baseline methods: In Section IV-D, we compare DCAPNet with several classical and state-of-the-art algorithms that have gained prominence in recent years. These comparative methods encompass eight traditional approaches: Tophat [5], MaxMedian [6], LCM [7], MPCM [8], PSTNN [13], and SRWS [14]. In addition, four deep learning-based methods are included in the comparison: ACM [26], DNANet [29], AGPCNet [19], and RPCANet [31].

Evaluation metrics: As we model the diminutive target detection task as a semantic segmentation issue, we employ classical metrics from semantic segmentation to assess the detection efficacy of the proposed DCAPNet on both datasets. These metrics include F-measure, mean intersection over union (mIoU), precision, and recall. F-measure and mIoU, respectively, reflect the ability of the network to balance precision and recall and describe target shapes, defined as follows:

$$ \mathrm{Fmeasure} = \frac{2 \times \mathrm{Precsion} \times \mathrm{Recall}}{\mathrm{Precsion} + \mathrm{Recall}} \tag{11} $$

$$ \mathrm{mIoU} = \frac{\#\,\text{Area of Overlap}}{\#\,\text{Area of Union}}. \tag{12} $$

Simultaneously, to achieve a more nuanced and equitable evaluation between deep learning models and traditional algorithms, we incorporate the specialized metric of normalized intersection over union (nIoU), specifically tailored for the assessment of infrared small target detection

$$ \mathrm{nIoU} = \frac{1}{N}\sum_{i}^{N} \frac{\mathrm{TP}[i]}{T[i] + P[i] - \mathrm{TP}[i]} \tag{13} $$

where TP, $T$, $P$, and $N$ represent true positive rate, true, positives, and the total number of samples, respectively. In addition, the receiver operating characteristic (ROC) curve is employed to depict the variation trend of the true positive rate (TPR) relative to different false positive rates (FPRs)

$$ \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP + FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP + TN}} \tag{14} $$

TABLE I
ABLATION STUDY OF EACH MODULE IN DCAPNET

| Backbone | DPC | SAF | SIRST-Aug | | | IRSTD-1k | | |
|---|---|---|---|---|---|---|---|---|
| | | | mIoU | F-measure | nIoU | mIoU | F-measure | nIoU |
| ResNet-18 | | | 0.6331 | 0.7753 | 0.6235 | 0.4112 | 0.5827 | 0.3671 |
| ResNet-18 | ✓ | | 0.6772 | 0.8075 | 0.6708 | 0.4684 | 0.6379 | 0.4216 |
| ResNet-18 | | ✓ | 0.7002 | 0.8237 | 0.7078 | 0.6036 | 0.7528 | 0.5915 |
| ResNet-18 | ✓ | ✓ | **0.7249** | **0.8405** | **0.7205** | **0.6437** | **0.7802** | **0.6541** |
| ResNet-34 | | | 0.6520 | 0.7894 | 0.6614 | 0.3979 | 0.5693 | 0.3608 |
| ResNet-34 | ✓ | | 0.6788 | 0.8087 | 0.6778 | 0.4381 | 0.5933 | 0.4075 |
| ResNet-34 | | ✓ | 0.6798 | 0.8093 | 0.7029 | 0.5692 | 0.7358 | 0.5569 |
| ResNet-34 | ✓ | ✓ | **0.6886** | **0.8156** | **0.7056** | **0.6112** | **0.7597** | **0.6016** |

TABLE II
ABLATION STUDY OF DCM MODULE SCALE

| DCM size | SIRST-Aug | | | IRSTD-1k | | |
|---|---|---|---|---|---|---|
| | mIoU | F-measure | nIoU | mIoU | F-measure | nIoU |
| (2) | 0.6952 | 0.8241 | 0.7064 | 0.6348 | 0.7766 | 0.6118 |
| (3) | 0.6949 | 0.8202 | 0.7071 | 0.6081 | 0.7563 | 0.6130 |
| (5) | 0.7008 | 0.8260 | 0.7089 | 0.6357 | 0.7793 | 0.6174 |
| (2,3) | 0.7043 | 0.8265 | 0.7109 | 0.6449 | 0.7838 | 0.6325 |
| (3,5) | 0.7036 | 0.8261 | 0.7093 | 0.6408 | 0.7811 | 0.6292 |
| (2,3,5) | 0.7168 | 0.8346 | 0.7140 | 0.6536 | 0.7885 | 0.6487 |
| (3,5,6) | 0.7066 | 0.8281 | 0.7123 | 0.6470 | 0.7854 | 0.6349 |
| (2,3,5,8) | **0.7297** | **0.8426** | **0.7195** | **0.6614** | **0.7970** | **0.6703** |
| (3,5,6,8) | **0.7280** | **0.8411** | **0.7147** | **0.6592** | **0.7896** | **0.6521** |

where TN, TP, FP, and FN represent true negatives, true positives, false positives, and false negatives, respectively. The area under the curve (AUC) is utilized as a quantitative metric to evaluate the detection capabilities of different models.

## C. Ablation Study

To scrutinize the rationality and efficacy of the constituent modules within DCAPNet, we conduct an ablation study by constructing variant networks with diverse module configurations.

*The effectiveness of backbone, DPC, and SAF:* we opted to compare the performance of pretrained ResNet-18 and ResNet-34 as backbone architectures to prevent the disappearance of target features in deeper layers. In addition, we incorporated DPC and SAF modules into each backbone network separately to validate the impact of the two modules. As presented in Table I, the results highlight the maximum values, denoted in red bold font, with the maximum values for each backbone indicated in black bold font. Analysis of the testing outcomes on two datasets elucidates that the inclusion of both DPC and SAF modules augments network detection performance, even under constant backbone architectures. Furthermore, the simultaneous addition of both modules attains optimal performance, underlining their synergistic impact on detection accuracy. Conversely, when confronting disparate backbone architectures, the deeper ResNet-34 backbone exhibits a propensity for suboptimal small target detection. This observation underscores the pivotal role of backbone depth in influencing detection efficacy, particularly evident in scenarios where small targets are at risk of being overshadowed within the intricate network hierarchy.

*The effect of DCM size in DPC:* In this investigation, we designate the internal kernel scale within the DCM as a pivotal parameter, aiming to assess the influence of various DCM scales on the model. Employing DPC modules equipped with DCM modules of varying scales, we conduct a comprehensive comparative analysis, meticulously evaluating the performance across diverse DCM kernel scale configurations, as shown in Table II. Table II showcases several distinct DCM kernel scale combinations, namely (2), (3), (5), (2,3), (3,5), (2,3,5), (3,5,6), (2,3,5,8), and (3,5,6,8). Each element within these groups corresponds to the scale of the filtering convolutional kernel in the DCM module, with the number of elements indicating the count of DCM modules. The outcomes presented in Table II illustrate that the incorporation of multiscale feature learning yields superior outcomes compared to single-scale approaches, attributing the superiority to the increased contextual information contained in multiscale features. In addition, a positive correlation is observed between the quantity of DCM modules within the DPC module and the concurrent enhancement in network performance. Strikingly, the apical performance is consistently realized in configurations featuring four distinct kernel sizes across

TABLE III
Ablation Study of SAF Module

| Backbone | SA | BAF | SIRST-Aug | | | IRSTD-1k | | |
|---|---|---|---|---|---|---|---|---|
| | | | mIoU | F-measure | nIoU | mIoU | F-measure | nIoU |
| ResNet-18 | | | 0.6667 | 0.8001 | 0.7001 | 0.5695 | 0.6138 | 0.5279 |
| ResNet-18 | ✓ | | 0.7070 | 0.8283 | 0.7047 | 0.6271 | 0.7708 | 0.6026 |
| ResNet-18 | | ✓ | 0.7099 | 0.8304 | 0.7022 | 0.6135 | 0.7604 | 0.6010 |
| ResNet-18 | ✓ | ✓ | **0.7182** | **0.8360** | **0.7044** | **0.6362** | **0.7776** | **0.6125** |

TABLE IV
Parameter Settings for the Compared Methods

| Methods | Parameter settings |
|---|---|
| Tophat | Structure size = 5×5 |
| MaxMedian | Patch size = 5×5 |
| LCM | Window size = $\{1,2,3,4\}$ |
| MPCM | Window size = $\{1,2,3,4\}$ |
| PSTNN | Patch size: 40×40, slide step: 40, $\lambda = 0.6/\sqrt{\max(n_1,n_2)\times n_3}$, $\varepsilon = 10^{-7}$ |
| SRWS | Patch size: 50×50, slide step: 10, $\beta = 1/\sqrt{\min{(m,n)}}$ |

both datasets. The maximum and second-maximum values in the table are respectively highlighted in red bold font and black bold font, providing a clear delineation of the most salient outcomes. This study illuminates the influence of DCM scales on network performance, underscoring the significance of multiscale feature learning and emphasizing the role of the DCM module configuration in optimizing the effect of neural networks.

*The effectiveness of SA block and BAF block in SAF:* In Table III, we present a meticulous evaluation of four distinct configurations: the singular integration of the SA module, the BAF module, the exclusion of both SA and BAF modules, and the proposed SAF module. These configurations are devised to elucidate the individual and combined impacts of SA and BAF blocks on network performance. The results in Table III underscore that both SA and BAF modules independently enhance the feature representation capability of the network. Notably, the findings reveal that the amalgamation of SA and BAF blocks within the SAF module yields superior detection performance. The superiority can be ascribed to the distinctive capability of SAF modules to not only integrate feature information within the same layer across spatial and channel dimensions but also fuse feature representations from higher and lower levels across both dimensions. The maximal values in the table are highlighted in red bold font. The study provides a comprehensive assessment of the distinct contributions of SA and BAF modules, emphasizing the synergistic benefits of their combination within the proposed SAF module.

## D. Comparison With the State of the Art

To further substantiate the detection capabilities of the proposed DCAPNet, we conducted a comprehensive evaluation through both quantitative and qualitative comparisons with the state-of-the-art methodologies on the IRSTD-1 k, SIRST-Aug, and NUDT-SIRST datasets. The parameters of the compared methods follow the settings specified in their respective papers, as delineated in Table IV.

1) Qualitative comparison: Figs. 7 and 8 illustrate the qualitative comparison results of our algorithm with the other ten advanced methodologies on the IRSTD-1 k, SIRST-Aug, and NUDT-SIRST datasets. In Fig. 7, the delineation of detection outcomes is symbolized by distinct bounding box colors: red signifies accurately detected target regions, yellow highlights regions indicative of false-positive identifications, and blue represents regions where targets were missed. The upper left region of the detection results is annotated with the respective detection method.

The meticulous examination of Figs. 7 and 8 reveals insights into the performance characteristics of various detection methodologies on the SIRST-Aug, IRSTD-1 k, and NUDT-SIRST datasets. Among the traditional techniques, the Tophat and MaxMedian methods based on background suppression exhibit susceptibility to noise interference, resulting in pronounced noise artifacts and background clutter in their respective detection outputs. LCM and MPCM relying on local contrast exhibit limited capabilities in suppressing complex backgrounds, leading to the elevated presence of clutter in the detection images and consequently generating a notable number of false positives. Although the two methods based on optimized PSTNN and SRWS have relatively few false positives, when the target and background are similar, both are significantly suppressed, resulting in missed detection. Critically, these traditional methods generally provide only approximate target localization without accurate segmentation of the target region. These traditional methods rely heavily on manually extracted features, which cannot completely filter out the complex changing background and cannot adapt to the change of target size, so it is easy to produce a large number of false alarm regions in complex scenes. In contrast, deep learning-based methods universally outperform traditional methods in detection performance. The

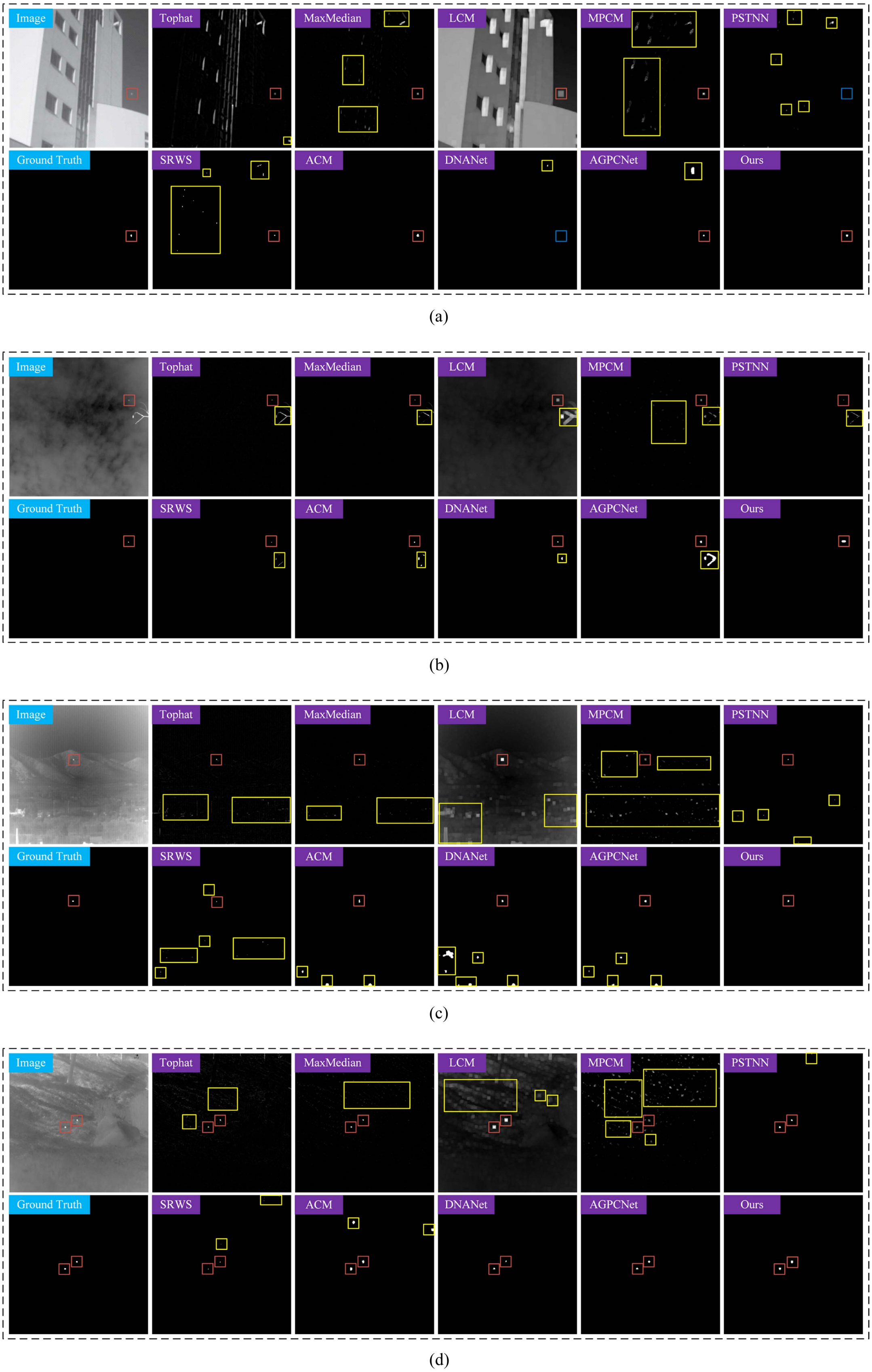
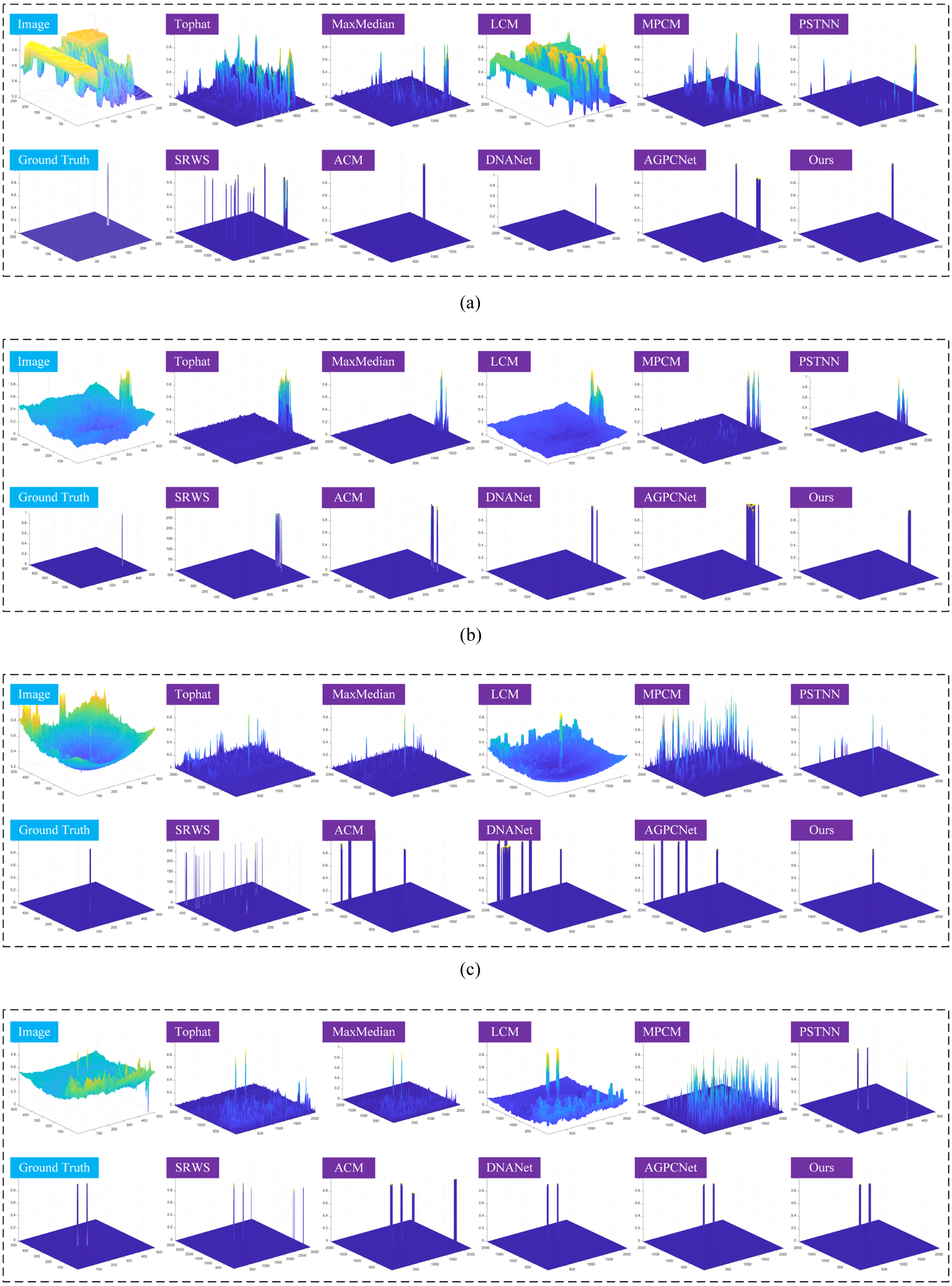


Fig. 7. Qualitative detection results of four infrared scenes by various detection methods. (a) Scene I. (b) Scene II. (c) Scene III. (d) Scene IV.



(a)

(b)

(c)

(d)

Fig. 8. 3-D visualization of qualitative results of four infrared scenes by various detection methods. (a) Scene I. (b) Scene II. (c) Scene III. (d) Scene IV.

superiority stems from the adaptability of deep learning methods to variations in complex backgrounds and target sizes present in both datasets, which is beyond the adaptability of manually designed features and prior knowledge incorporated into traditional methods. However, deep learning methods such as ACM tend to produce false positives due to their only fusion of details and high-level semantics, ignoring potential semantic conflicts arising during the fusion process. DNANet, AGPCNet, and RPCANet exhibit some false alarm regions due to not considering the multiscale feature representation and the information loss problem it generates. Conspicuously, our proposed DCAPNet model demonstrates superior precision in target detection and

TABLE V
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS

| Methods | SIRST-Aug | | | | IRSTD-1k | | | | NUDT-SIRST | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mIoU | F-measure | nIoU | AUC | mIoU | F-measure | nIoU | AUC | mIoU | F-measure | nIoU | AUC |
| Tophat [5] | 0.0051 | 0.0039 | 0.0101 | 0.6832 | 0.0550 | 0.0713 | 0.1043 | 0.7775 | 0.0287 | 0.0557 | 0.0093 | 0.9321 |
| MaxMedian [6] | 0.0032 | 0.0029 | 0.0063 | 0.6366 | 0.0691 | 0.0619 | 0.1293 | 0.7175 | 0.0101 | 0.0201 | 0.0036 | 0.7894 |
| LCM [7] | 0.0051 | 0.1210 | 0.0102 | 0.9186 | 0.0008 | 0.1220 | 0.1015 | **0.9737** | 0.0012 | 0.0024 | 0.0147 | 0.9298 |
| MPCM [8] | 0.2686 | 0.1438 | 0.4235 | 0.7822 | 0.0692 | 0.1114 | 0.1294 | 0.8063 | 0.0656 | 0.1231 | 0.1782 | 0.8365 |
| PSTNN [13] | 0.4163 | 0.2027 | 0.5879 | 0.6365 | 0.3537 | 0.2718 | 0.5226 | 0.7352 | 0.2697 | 0.4248 | 0.2073 | 0.7278 |
| SRWS [14] | 0.5041 | 0.3826 | 0.6195 | 0.7208 | 0.4950 | 0.3817 | 0.5903 | 0.7946 | 0.2958 | 0.5083 | 0.3192 | 0.8901 |
| ACM [26] | 0.7058 | 0.8276 | 0.7012 | **0.9497** | 0.4507 | 0.6213 | 0.4058 | 0.9290 | 0.6373 | 0.7785 | 0.6451 | 0.9753 |
| DNANet [29] | **0.7252** | 0.8407 | 0.6972 | 0.9286 | 0.6532 | **0.7902** | **0.6665** | 0.9017 | **0.9050** | **0.9501** | **0.9176** | **0.9825** |
| AGPCNet [19] | 0.7106 | 0.8309 | 0.7108 | 0.9312 | **0.6578** | 0.7138 | 0.6323 | 0.9114 | 0.8780 | 0.9350 | 0.9020 | 0.9738 |
| RPCANet [31] | 0.7162 | **0.8471** | **0.7133** | 0.8906 | 0.6182 | 0.7640 | 0.6163 | 0.8798 | 0.8767 | 0.9343 | 0.9160 | 0.8714 |
| DCAPNet (Ours) | **0.7454** | **0.8541** | **0.7203** | **0.9621** | **0.6681** | **0.8017** | **0.6964** | **0.9402** | **0.9274** | **0.9623** | **0.9416** | **0.9876** |

segmentation, with minimal occurrences of missed detections and false positives. The heightened performance is attributed to the adaptive multiscale feature learning facilitated by the DPC module and the SAF module, enabling DCAPNet to adeptly accommodate variations in target shape, size, and diverse complex scene categories.

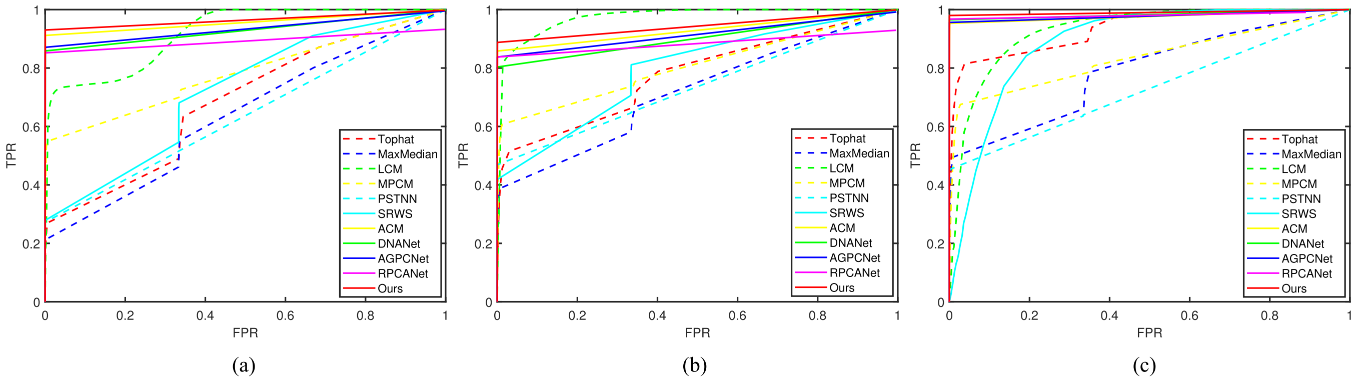
2) Quantitative comparison: For a precise evaluation of the detection performance of DCAPNet, we employ numerical methods for objective evaluation. Table V presents the results of the quantitative comparisons between DCAPNet and ten other advanced detection methods. The maximum value for each metric is highlighted in red bold font, while the second-maximum value is indicated in blue bold font. It is noteworthy that one max-pooling operation in the first downsampling stage was removed due to discrepancies in the input image size of the original ACM (480×480) compared to the size discussed in Section IV-B of this article. Remaining consistent with the parameters of the original paper, we retrained four other CNN-based detection algorithms using the datasets and input image size specified in this study.

It is evident from Table V that our proposed DCAPNet model demonstrates the best performance across two datasets, achieving the highest metrics of mIoU, nIoU, and F-measure at 0.9274, 0.9416, and 0.9623, respectively. In contrast to conventional algorithms, our method exhibits a significant enhancement in performance metrics. Due to the challenging scenarios present in the SIRST-Aug, IRSTD-1 k, and NUDT-SIRST datasets, including diverse target sizes, varying background complexities, and dynamic signal-to-noise ratios. DCAPNet effectively leverages the advantages of deep learning algorithms by learning highly discriminative semantic features from diverse training data, thereby achieving robust target detection results. Traditional algorithms, such as LCM and PSTNN, heavily rely on prior knowledge and necessitate manual parameter adjustments to adapt to diverse scenarios. Notably, while LCM achieves a high AUC, reaching 0.9737 on the IRSTD-1 k dataset, its performance across the other three metrics is lower. The disparity indicates that while LCM attains a heightened detection rate, it also encounters a substantial FPR, leading to a relatively high AUC but lower detection accuracy.

The DCAPNet model consistently exhibits superior performance compared to four other CNN-based methods. The achievement can be ascribed to the proposed DPC module and SAF module, which effectively preserve and enhance the feature responses of targets within the deep layers, thereby contributing to the improved detection performance of the model. Primarily, the network conducts dynamic multiscale feature learning at the deepest layer of feature extraction, reducing the loss of contextual information caused by pooling operations and better-capturing information at multiple feature scales. Subsequently, attention modulation and fusion mechanisms are applied to the low-level details and deep-level semantics of small targets, preventing the submergence of crucial target features within the conflicting information. In essence, the prowess emanates of DCAPNet from its holistic integration of dynamic multiscale feature learning and stochastic attention modulation techniques, collectively fortifying its capacity to discern and highlight salient features of infrared small targets amidst deep-layer network responses.

We plot the ROC curves for various methods, utilizing TPR and FPR, respectively, represented on the vertical and horizontal axes, as illustrated in Fig. 9. Combining the AUC values with the ROC curves reveals that, as FPR increases, DCAPNet consistently maintains the highest detection probability and AUC values. The observation underscores the proposed model effectively suppresses background information, mitigates noise interference, and exhibits optimal comprehensive detection capabilities.

To further validate the detection efficiency of the proposed method, we calculate the parameter quantity and inference time for ten detection methods, as shown in Table VI. Although our designed model shows the best performance on other evaluation metrics, the network parameters increase by 5.15 M compared to the AGPCNet model. The reason is that our proposed DPC module combines four DCM modules, integrating multiscale information of images, enhancing the feature representation of small targets, but also increasing the complexity of the model. It can be seen from the table that Tophat and MaxMedian require the least inference time among traditional algorithms, but



(a) (b) (c)

Fig. 9. ROC curves of 11 infrared small target detection methods. (a) SIRST-Aug dataset. (b) IRSTD-1 k dataset. (c) NUDT-SIRST dataset.

TABLE VI
PARAMETER QUANTITY AND INFERENCE TIME FOR 11 DETECTION METHODS

| Methods | Params (M) | Times (s) |
|---|---|---|
| Tophat [5] | – | 0.0061 |
| MaxMedian [6] | – | 0.0075 |
| LCM [7] | – | 20.8976 |
| MPCM [8] | – | 0.0429 |
| PSTNN [13] | – | 0.2194 |
| SRWS [14] | – | 1.3358 |
| ACM [26] | 0.52 | 0.0718 |
| DNANet [29] | 4.70 | 0.3391 |
| AGPCNet [19] | 12.36 | 0.2712 |
| RPCANet [31] | 0.68 | 0.0964 |
| DCAPNet (ours) | 17.51 | 0.2693 |

their detection performance is relatively poor, and they cannot adapt to complex environments. Among deep learning-based algorithms, the inference time of our method is second only to the ACM network, which can process images in real-time while maintaining better detection performance.

## V. Discussion

In summation, the proposed model in this article exhibits heightened capabilities in both target detection and segmentation. Initially, elaborate ablation experiments were conducted to validate the soundness and efficacy of the introduced modules, encompassing the role of the backbone, the selection of DCM scales, and the impact of SA attention and BAF modules. Subsequently, we compared DCAPNet with ten other advanced detection algorithms, conducting qualitative and quantitative analyses of the experimental outcomes from both subjective and objective perspectives, respectively. Qualitative experimental results underscore the limitations of traditional detection algorithms, which incorporate diverse prior knowledge such as structural tensors, local contrast, and regularization terms. These methods exhibit limited adaptability to variations in targets and scenes, resulting in a notable propensity for generating false positives and missed detection. Concurrently, the mIoU, nIoU, and F-measure metrics of traditional methods differed significantly from those based on deep learning, as delineated in Table V. The discrepancy is attributed to the fact that classic traditional detection algorithms primarily focus on locating targets rather than precisely segmenting them. Consequently, owing to their heavy reliance on manually tuned parameters, traditional methods are susceptible to noise interference, thereby constraining their generalization capability and exhibiting inferior detection performance in contrast to their deep learning counterparts.

Within the realm of deep learning-based methods, the ACM approach merely integrates detailed information and high-level semantics, neglecting the distinctions between targets and backgrounds and lacking focused attention on target regions. The DNANet model exhibits a complex design but fails to exploit multiscale contextual information effectively. AGPCNe generates a considerable amount of conflicting information during the feature fusion process, potentially resulting in the loss of details related to target features. RPCANet may lead to missed detections in small and complex situations. The diverse experimental results presented in Section IV-D collectively substantiate the robustness of the proposed DCAPNet model across various complex backgrounds and its more accurate detection capabilities.

Nevertheless, the persisting challenge lies in the circumstance where small targets occupy only a limited number of pixels and wield relatively minor weights in the loss function, further research is imperative in future work to delve deeper into this aspect. As shown in Table VI, our proposed module increases the number of network parameters, making our network more network parameters than other networks. Therefore, we will continue to explore how to reduce model complexity and refine the application of attention mechanisms within the proposed framework.

## VI. Conclusion

To enhance the precision of small target detection in infrared images, this article introduces a novel approach termed the DCAPNet designed specifically for detecting weak small targets in the infrared spectrum. To tackle the issue of target loss in deep networks, we formulate a DPC module that assimilates multiple DCM modules. The construction enables the adaptive learning of multiscale features related to targets, thereby effectively augmenting the context information for small targets. Subsequently, we design the SAF module to extract and apply interpixel correlation information within the same layer features, preventing target information from being submerged in semantic conflict during feature fusion. The module fully incorporates the shallow positional information and deep semantic features of the target, accentuating the target features in different layers while concurrently suppressing extraneous background information. In addition, the impact of each module is verified by ablation studies, and objective quantitative as well as subjective qualitative comparisons are conducted with ten other detection approaches. Across the SIRST-Aug, IRSTD-1 k, and NUDT-SIRST datasets, DCAPNet consistently exhibits superior performance compared to other methods, manifesting heightened robustness in the face of diverse and complex backgrounds. The methods based on graph signal processing and graph neural network require less labeled data [54], [55], [56], [57], and can be applied to infrared small target detection in the future to further reduce the dependence of the model on the amount of data.

## References

[1] Z. Wu, N. Fuller, D. Theriault, and M. Betke, "A thermal infrared video benchmark for visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 201–208.

[2] D. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabally, and C. Quek, "Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 1993–2016, Aug. 2017.

[3] S. S. Rawat, S. K. Verma, and Y. Kumar, "Review on recent development in infrared small target detection algorithms," *Procedia Comput. Sci.*, vol. 167, pp. 2496–2505, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920307687

[4] H. Wang, M. Shi, and H. Li, "Infrared dim and small target detection based on two-stage u-skip context aggregation network with a missed-detection-and-false-alarm combination loss," *Multimedia Tools Appl.*, vol. 79, pp. 35383–35404, 2020.

[5] V. T. Tom, T. Peli, M. Leung, and J. E. Bondaryk, "Morphology-based algorithm for point target detection in infrared backgrounds," in *Proc. Defense, Secur., Sens.*, vol. 1954, 1993, pp. 2–11. [Online]. Available: https://api.semanticscholar.org/CorpusID:119943629

[6] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," in *Proc. Opt. Photon.*, vol. 3809, 1999pp. 74–83. [Online]. Available: https://api.semanticscholar.org/CorpusID:129888940

[7] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.

[8] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320316300358

[9] 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.

[10] 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.

[11] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.

[12] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017.

[13] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 382. [Online]. Available: https://www.mdpi.com/2072-4292/11/4/382

[14] T. Zhang, Z. Peng, H. Wu, Y. He, C. Li, and C. Yang, "Infrared small target detection via self-regularized weighted sparse model," *Neurocomputing*, vol. 420, pp. 124–148, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220313461

[15] J. Wang, W. Li, M. Zhang, and J. Chanussot, "Large kernel sparse ConvNet weighted by multi-frequency attention for remote sensing scene understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 16, 2023, Art. no. 5626112.

[16] J. Wang, M. Zhang, W. Li, and R. Tao, "A multistage information complementary fusion network based on flexible-mixup for HSI-X image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 14, 2023, doi: 10.1109/TNNLS.2023.3300903.

[17] Y. Gao, W. Li, J. Wang, M. Zhang, and R. Tao, "Relationship learning from multisource images via spatial-spectral perception network," *IEEE Trans. Image Process.*, vol. 33, pp. 3271–3284, May 2, 2024.

[18] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9813–9824, Nov. 2021.

[19] T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-guided pyramid context networks for detecting infrared small target under complex background," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 4, pp. 4250–4261, Aug. 2023.

[20] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proc. 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 867–876.

[21] M. Liu, H. yuan Du, Y. jin Zhao, L. quan Dong, and M. Hui, *Image Small Target Detection Based on Deep Learn. With SNR Controlled Sample Gener.*, Warsaw, Poland: De Gruyter Open Poland, 2022, pp. 211–220. [Online]. Available: https://doi.org/10.1515/9783110584974-025

[22] 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.

[23] M. Zhang, K. Yue, J. Zhang, Y. Li, and X. Gao, "Exploring feature compensation and cross-level correlation for infrared small target detection," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1857–1865. [Online]. Available: https://api.semanticscholar.org/CorpusID:252782178

[24] M. Zhang, H. Yang, K. Yue, X. Zhang, Y. Zhu, and Y. Li, "Thermodynamics-inspired multi-feature network for infrared small target detection," *Remote Sens.*, vol. 15, no. 19, 2023, Art. no. 4716. [Online]. Available: https://www.mdpi.com/2072-4292/15/19/4716

[25] M. Zhang et al., "RKformer: Runge-Kutta transformer with random-connection attention for infrared small target detection," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1730–1738. [Online]. Available: https://api.semanticscholar.org/CorpusID:252782766

[26] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 949–958.

[27] M. Zhang, H. Yang, J. Guo, Y. Li, X. Gao, and J. Zhang, "IRPruneDet: Efficient infrared small target detection via wavelet structure-regularized soft channel pruning," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 7224–7232.

[28] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002013.

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

[30] B. Yang, X. Zhang, J. Zhang, J. Luo, M. Zhou, and Y. Pi, "EFLNet: Enhancing feature learning network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5906511.

[31] 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.*, 2024, pp. 4797–4806.

[32] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE ICASSP Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 2235–2239.

[33] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3080–3089.

[34] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[35] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[36] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, *arXiv:1706.05587*. [Online]. Available: https://api.semanticscholar.org/CorpusID:22655199

[37] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818. [Online]. Available: https://api.semanticscholar.org/CorpusID:3638670

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6230–6239.

[39] P. Hu and D. Ramanan, “Finding tiny faces,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1522–1530.

[40] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *Proc. 36th Int. Conf. Mach. Learn.*, (Proceedings of Machine Learning Research Series), K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, Jun. 2019, pp. 7354–7363. [Online]. Available: https://proceedings.mlr.press/v97/zhang19d.html

[41] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[42] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. conf. Comput. Vis.*, Berlin, Germany: Springer-Verlag, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[44] B. Bosquet, M. Mucientes, and V. M. Brea, “STDnet: A convnet for small target detection,” in *Proc. Brit. Mach. Vis. Conf.*, 2018, Art. no. 253. [Online]. Available: https://api.semanticscholar.org/CorpusID:52287549

[45] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, and Z. Han, “Effective fusion factor in FPN for tiny object detection,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1159–1167.

[46] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[47] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer Int. Publishing, 2015, pp. 234–241.

[48] M. Zhao, L. Cheng, X. Yang, P. Feng, L. Liu, and N. Wu, “TBC-Net: A real-time detector for infrared small target detection using semantic constraint,” 2019, *arXiv:2001.05852*. [Online]. Available: https://api.semanticscholar.org/CorpusID:210701230

[49] Y. Wu and K. He, “Group normalization,” *Int. J. Comput. Vis.*, vol. 128, pp. 742–755, 2020.

[50] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “ShuffleNet v2: Practical guidelines for efficient CNN architecture design,” in *Proc. Eur. Conf. Comput. Vis.*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer Int. Publishing, 2018, pp. 122–138.

[51] M. Lin, Q. Chen, and S. Yan, “Network in network,” 2013, *arXiv:1312.4400*. [Online]. Available: https://api.semanticscholar.org/CorpusID:16636683

[52] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[53] M. A. Rahman et al., “Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation,” in *Advances in Visual Computing*, G. Bebis et al., Eds. Cham, Switzerland: Springer Int. Publishing, 2016, pp. 234–244.

[54] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, “Graph signal processing: Overview, challenges, and applications,” *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[55] G. Leus, A. G. Marques, J. M. Moura, A. Ortega, and D. I. Shuman, “Graph signal processing: History, development, impact, and outlook,” *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 49–60, Jun. 2023.

[56] S. M. J. S. K. Giraldo, H. Jhony, and T. Bouwmans, “The emerging field of graph signal processing for moving object segmentation,” in *Frontiers of Computer Vision*, H. Jeong and K. Sumi, Eds. Cham, Switzerland: Springer Int. Publishing, 2021, pp. 31–45.

[57] G. Cheung, E. Magli, Y. Tanaka, and M. K. Ng, “Graph spectral image processing,” *Proc. IEEE*, vol. 106, no. 5, pp. 907–930, May 2018.

**Xiaolong Chen** received the B.S. degree in mechanical and electrical engineering from the Wuhan University of Technology, Wuhan, China, in 2020. He is currently working toward the doctor’s degree in optical engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, Changchun, China.

His research interests include object detection and semantic segmentation.

**Jing Li** received the graduation degree in control engineering from Naval Aviation University, Yantai, China.

She is currently with the Beijing Institute of Tracking and Communication Technology, Beijing, China. Her research focuses on optical detection.

**Tan Gao** received the B.S. degree in mechatronic engineering from Hainan University, Hainan, China, in 2019. He is currently working toward the doctor’s degree in mechatronic engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

His research interests include image processing and imaging electronic.

**Yongjie Piao** received the B.S. degree in electronic information science and technology from Jilin University, Changchun, China, in 2006, and the Ph.D. degree in optical engineering from the Changchun Institute of Optics and Mechanics, Chinese Academy of Sciences, in 2011.

He is currently with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests include on-board electronics systems and image processing.

**Haolin Ji** received the B.S. degree in mechatronic engineering from the Beijing Institute of Technology, Beijing, China, in 2020. He is currently working toward the M.S. degree in mechatronic engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

His research interests include remote sensing image processing and satellite electronics.

**Biao Yang** received the B.S. degree in mechatronic engineering from Hainan University, Hainan, China, in 2019. He is currently working toward the M.S. degree in mechatronic engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

His research focuses on low-illumination image enhancement.

**Wei Xu** received the B.S. degree in mechatronic engineering from Jilin University, Changchun, China, in 2003, and the Ph.D. degree in mechanical and electronic engineering from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2008.

He is currently a Research Fellow and a Doctoral Supervisor in the field of mechatronic engineering. After graduation, he was with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include the integration technology of satellites and payloads and the high reliable electronic systems for aerospace.