Dense Nested Attention Network for Infrared Small Target Detection

Boyang Li*, Chao Xiao*, Longguang Wang*, Yingqian Wang*, Zaiping Lin*, Miao Li, Wei An*, and Yulan Guo*, Senior Member, IEEE

Abstract—Single-frame infrared small target (SIRST) detection aims at separating small targets from clutter backgrounds. With the advances of deep learning, CNN-based methods have yielded promising results in generic object detection due to their powerful modeling capability. However, existing CNN-based methods cannot be directly applied to infrared small targets since pooling layers in their networks could lead to the loss of targets in deep layers. To handle this problem, we propose a dense nested attention network (DNA-Net) in this paper. Specifically, we design a dense nested interactive module (DNIM) to achieve progressive interaction among high-level and low-level features. With the repetitive interaction in DNIM, the information of infrared small targets in deep layers can be maintained. Based on DNIM, we further propose a cascaded channel and spatial attention module (CSAM) to adaptively enhance multi-level features. With our DNA-Net, contextual information of small targets can be well incorporated and fully exploited by repetitive fusion and enhancement. Moreover, we develop an infrared small target dataset (namely, NUDT-SIRST) and propose a set of evaluation metrics to conduct comprehensive performance evaluation. Experiments on both public and our self-developed datasets demonstrate the effectiveness of our method. Compared to other state-of-the-art methods, our method achieves better performance in terms of probability of detection ($P_D$), false-alarm rate ($F_A$), and intersection of union ($IoU$).

Index Terms—Infrared small target detection, deep learning, dense nested interactive module, channel and spatial attention, dataset.

I. INTRODUCTION

SINGLE-FRAME infrared small target (SIRST) detection is widely used in many applications such as maritime surveillance [1], [2], early warning systems [3], [4], and precise guidance [5]. Compared to generic object detection, infrared small target detection has several unique characteristics: 1) **Small**: Due to the long imaging distance, infrared targets are generally small, ranging from one pixel to tens of pixels in the images. 2) **Dim**: Infrared targets usually have low signal-to-clutter ratio (SCR) and are easily immersed in heavy noise and clutter background. 3) **Shapeless**: Infrared small targets have limited shape characteristics. 4) **Changeable**: The sizes and shapes of infrared targets vary a lot among different scenarios.

To detect infrared small targets, numerous traditional methods have been proposed, including filtering-based methods [6], [9], local-contrast-based methods [10], [11], [12], [13], [14], [15], and low-rank-based methods [7], [8], [16], [17], [18], [19]. However, these traditional methods heavily rely on handcrafted features. Considering the characteristics of real scenes (e.g., target size, target shape, SCR, and clutter background) change dramatically, it is difficult to use handcrafted features and fixed hyper-parameters to handle such variations.

Different from traditional methods, CNN-based methods can learn features of infrared small targets in a data-driven manner. Liu et al. [20] proposed the first CNN-based SIRST detection method. They designed a multi-layer perceptron (MLP) network with 5 layers for infrared small target detection. Then, McIntosh et al. [21] fine-tuned several existing generic object detection networks (e.g., Faster-RCNN [22] and Yolo-v3 [23]) for infrared small target detection. Specifically, Dai et al. [24] proposed the first segmentation-based SIRST detection method. They designed an asymmetric contextual module (ACM) to replace the plain skip connection of Unet [25]. Although recent CNN-based methods have achieved
The state-of-the-art performance, most of them only fine-tuned these networks designed for generic objects. Since the size of infrared small targets is much smaller than generic objects, directly applying these methods for SIRST detection can easily lead to the loss of small targets in deep layers.

Inspired by the success of nested structure in medical image segmentation [26], [27], [28], [29] and hybrid attention in generic object detection [30], we propose a dense nested attention network (namely, DNA-Net) to maintain small targets in deep layers. Specifically, we design a tri-directional dense nested interactive module (DNIM) with a cascaded channel and spatial attention module (CSAM) to achieve progressive feature interaction and adaptive feature enhancement. Within our DNIM, multiple nodes are imposed on the pathway between the encoder and decoder sub-networks. As shown in Fig. 2(b), all nodes in our network are connected with each other to form a nested-shape network. Using DNIM, those middle nodes can receive features from their own and the adjacent two layers, leading to repetitive multi-layer feature fusion at deep layers. Through repetitive feature fusion and enhancement, our network can maintain the targets in deep layers. Meanwhile, contextual information of maintained small targets can be well incorporated and fully exploited. Otherwise, as shown in Fig. 2(a), the traditional U-shape network suffers from the loss of small targets in deep layers, which ultimately leads to inferior performance. In addition, we develop a novel infrared small target dataset (namely, NUDT-SIRST) to evaluate the performance of SIRST detection methods under different clutter backgrounds, target shapes, and target sizes. In summary, the contributions of this paper can be summarized as follows.

- We propose a DNA-Net to maintain small targets in deep layers. The contextual information of small targets can be well incorporated and fully exploited by repetitive feature fusion and enhancement.
- A dense nested interactive module and a channel-spatial attention module are proposed to achieve progressive feature fusion and adaptive feature enhancement.
- We develop an infrared small target dataset (namely, NUDT-SIRST). To the best of our knowledge, our dataset is the largest dataset with numerous categories of target shapes, various target sizes, diverse clutter backgrounds, and ground truth annotations.
- Experiments on both public and our NUDT datasets demonstrate the superior performance of our method. Compared to existing methods, our method is more robust to the variations of clutter background, target size, and target shape (as shown in Fig. 1).

This paper is organized as follows: In Section II, we briefly review the related work. In Section III, we introduce the architecture of our DNA-Net and our self-developed dataset in details. In Section IV, we introduce our self-developed NUDT-SIRST dataset in details. The experimental results are represented in Section V. Section VI gives the conclusion.

II. RELATED WORK

In this section, we briefly review the major works in SIRST detection and corresponding datasets.

A. Single-Frame Infrared Small Target Detection

SIRST detection has been extensively investigated for decades. The traditional paradigm achieves SIRST detection by measuring the discontinuity between targets and backgrounds. Typical methods include filtering-based methods [6], [9], local contrast measure based methods [10], [11], [12], [13], [14], [15], and low rank based methods [7], [8], [16], [17], [18], [19]. Considering real scenes are much more complex with dramatic changes target size, shape, and clutter background, it is difficult to use handcrafted features and fixed hyper-parameters to handle such variations. To address this problem, recent CNN-based methods learn trainable features in a data-driven manner. Thanks to the large quantity of data and the powerful model fitting capability of CNNs, these methods achieve better performance than traditional ones.

Existing CNN-based methods can be divided into detection based methods and segmentation based methods. Liu et al. [20] first introduced a generic target detection framework for infrared small target detection. They designed a multi-layer perception (MLP) network with 5 layers for infrared small target detection. Then, McIntosh et al. [21] fine-tuned several generic target detection network (e.g., Faster-RCNN [22] and Yolo-v3 [23]) and used the optimized eigen-vectors as input to achieve improved performance.

Recently, segmentation-based methods have attracted increasing attention. That is because, these methods can produce both pixel-level classification and localization outputs. Dai et al. [24] proposed the first segmentation-based network (i.e., ACM). They designed an asymmetric contextual module to aggregate features from shallow layers and deep layers. Then, Dai et al. [31] further improved their ACM by introducing a dilated local contrast measure. Specifically, a feature cyclic shift scheme was designed to achieve a trainable local contrast measure. Moreover, Wang et al. [32] decomposed
the infrared target detection problem into two opposed subproblems (i.e., miss detection and false alarm) and used a conditional generative adversarial network (CGAN) to achieve the trade-off between miss detection and false alarm for infrared small target detection.

Although the performance is continuously improved by recent networks, the loss of small targets in deep layers still remains. This problem ultimately results in the poor robustness to dramatic scene changes (e.g., clutter background, targets with different SCR, shape, and size).

B. Datasets for SIRST Detection

Existing open-source dataset in infrared small target detection is scarce, most traditional methods are evaluated on their in-house datasets. Only a few infrared small target datasets are released by CNN-based methods [24], [32]. Wang et al. [32] built the first big and open SIRST dataset. This dataset includes 10000 training images and 100 test images. However, many targets in this dataset do not meet the definition of society of photo-optical instrumentation engineers (SPIE) [33] and have obvious synthesized traces with illogical annotations. These problems may lead to the inapplicability toward SIRST detection. Dai et al. [24] built the first real SIRST dataset with high-quality images and labels. However, the number of images in NUAA-SIRST is 427 (256 for training), which cannot well cover dramatic scene changes in infrared small target detection. Moreover, these real infrared data are all manually labelled with many inaccurately labeled pixels.

Although these open-sourced datasets greatly prompt the prosperity of SIRST detection, their limited data capacity, data variety, and poor annotation hinder the further development of this field. Synthesized data can be easily generated to achieve higher variety and annotation quality at very low cost (i.e., time and money). Hence, we developed a new NUDT-SIRST dataset with numerous categories of target, various target sizes, diverse clutter backgrounds, and accurate annotations. The superiority of our dataset is evaluated in Section V.

III. METHODOLOGY

In this section, we introduce our DNA-Net in details.

A. Overall Architecture

As illustrated in Fig. 3, our DNA-Net takes a SIRST image as its input and sequentially performs feature extraction (Section III-B), feature pyramid fusion (Section III-C), and eight-connected neighborhood clustering (Section III-D) to generate the detection results.

Section III-B introduces the motivation of our feature extraction module and the architecture of the dense nested interactive module (DNIM) and the channel-spatial attention module (CSAM). Input images are first preprocessed and fed into the backbone of DNIM to extract multi-layer features. Then, multi-layer features are repetitively fused at the middle convolution nodes of skip connection and then are gradually passed into the decoder subnetworks. Due to the semantic gap at multi-layer feature fusion stage of DNIM, we used CSAM to adaptively enhance these multi-level features for achieving better feature fusion. Section III-C presents the feature pyramid fusion module. Enhanced multi-layer features at each scale are upsampled to the same size. Next, the shallow-layer features with rich spatial information and deep-layer features with high-level information are concatenated to generate robust feature maps. Section III-D elaborates the eight-connected neighborhood clustering module. Feature maps are fed into this module to calculate the spatial location of target centroid, which is then used for comparison in Section V.

B. The Feature Extraction Module

1) Motivation: As shown in Fig. 4(a), traditional U-shape structure [25] consists of an encoder, a decoder, and plain skip
connections. The encoder is used to enlarge the receptive field and extract high-level information. Decoder helps to recover the size of feature maps (which finally reach the same size as the input images) and achieve progressive multi-scale feature fusion. The plain skip connection acts as a bridge to pass these low-level and high-level features from encoder to decoder subnetworks.

To achieve powerful contextual information modeling capability, a straightforward way is to continuously increase the number of layers. In this way, high-level information can be obtained and larger receptive field can be achieved. However, infrared small targets are significantly different in their sizes, ranging from one pixel (i.e., point targets) to tens of pixels (i.e., extended targets). With the increase of network layers, high-level information of extended targets is obtained, while the point targets are easily lost after multiple pooling operation. Therefore, we should design a special module to extract high-level features and maintain the representation of small targets in the deep layers.

2) The Dense Nested Interactive Module: As shown in Fig. 4(b), we stack multiple U-shape sub-networks together to build a dense nested structure. Since the optimal receptive field for different sizes of targets varies a lot, these U-shape sub-networks with different depths are naturally suitable for targets with different sizes. Based on this idea, we impose sub-networks with different depths are naturally suitable for targets with different sizes. Based on this idea, we impose sub-networks with different depths are naturally suitable for targets with different sizes. Based on this idea, we impose sub-networks with different depths are naturally suitable for targets with different sizes.

The insight comes from the multiple U-shape sub-network stacking. As shown in Fig. 4(c) and (d), each node can receive features from its sub-networks. All of these middle nodes are densely connected with each other to form a nested-shape network. As shown in Fig. 4(c) and (d), each node can receive features from its own and the adjacent layers, leading to repetitive multi-layer feature fusion. As a result, the representations of small targets are maintained in the deep layers and thus better results can be achieved.

In this paper, we stack J layers of DNIM to form our feature extraction module. Without loss of generality, we take the $i^{th}$ ($i = 0, 1, 2, \ldots, J$) DNIM layer as an example to introduce this structure, as shown in Fig. 4(c) and (d). Assume $L^{i,j}$ denote the output of node $L^{i,j}$, where $i$ is the $i^{th}$ down-sampling layer along the encoder and $j$ is the $j^{th}$ convolutional layer of dense block along the plain skip pathway. When $j = 0$, each node only receives features from dense plain skip connection. The stack of feature maps represented by $L^{i,j}$ is computed as:

$$L^{i,j} = \mathcal{P}_{\text{max}}(\mathcal{F}(L^{i-1,j})), \quad (1)$$

where $\mathcal{F}(\cdot)$ denotes multiple cascaded convolution layers of the same convolution block. $\mathcal{P}_{\text{max}}(\cdot)$ denotes max-pooling with a stride of 2. When $j > 0$, each node receives outputs from three directions including dense plain skip connection and nested bi-direction interactive skip connection, the stack of feature maps represented by $L^{i,j}$ is generated as:

$$L^{i,j} = \mathcal{F}(L^{i,j}) = \mathcal{P}_{\text{max}}(\mathcal{F}(L^{i+1,j-1})) \cup \mathcal{U}(\mathcal{F}(L^{i-1,j})), \quad (2)$$

where $\mathcal{U}(\cdot)$ denotes the up-sampling layer, and $[\cdot, \cdot]$ denotes the concatenation layer.

3) Channel and Spatial Attention Module: As shown in Fig. 5, CSAM is used to reduce the semantic gap at the multi-layer feature fusion stage in DNIM.

The CSAM consists of two cascaded attention units. The feature maps $L^{i,j}$ from node $L^{i,j}$ ($i \in \{0, 1, 2, \ldots, I\}$, $j \in \{0, 1, 2, \ldots, J\}$) are sequentially processed by a 1D channel attention process can be computed as:

$$L^{i,j} = \mathcal{P}_{\text{max}}(\mathcal{F}(L^{i,j})), \quad (1)$$

where $\mathcal{F}(\cdot)$ denotes multiple cascaded convolution layers of the same convolution block. $\mathcal{P}_{\text{max}}(\cdot)$ denotes max-pooling with a stride of 2. When $j > 0$, each node receives outputs from three directions including dense plain skip connection and nested bi-direction interactive skip connection, the stack of feature maps represented by $L^{i,j}$ is generated as:

$$L^{i,j} = \mathcal{F}(L^{i,j}) = \mathcal{P}_{\text{max}}(\mathcal{F}(L^{i+1,j-1})) \cup \mathcal{U}(\mathcal{F}(L^{i-1,j})), \quad (2)$$

where $\mathcal{U}(\cdot)$ denotes the up-sampling layer, and $[\cdot, \cdot]$ denotes the concatenation layer.

The CSAM consists of two cascaded attention units. The feature maps $L^{i,j}$ from node $L^{i,j}$ ($i \in \{0, 1, 2, \ldots, I\}$, $j \in \{0, 1, 2, \ldots, J\}$) are sequentially processed by a 1D channel attention process can be computed as:

$$L^{i,j} = \mathcal{P}_{\text{max}}(\mathcal{F}(L^{i,j})), \quad (1)$$

where $\mathcal{F}(\cdot)$ denotes multiple cascaded convolution layers of the same convolution block. $\mathcal{P}_{\text{max}}(\cdot)$ denotes max-pooling with a stride of 2. When $j > 0$, each node receives outputs from three directions including dense plain skip connection and nested bi-direction interactive skip connection, the stack of feature maps represented by $L^{i,j}$ is generated as:

$$L^{i,j} = \mathcal{F}(L^{i,j}) = \mathcal{P}_{\text{max}}(\mathcal{F}(L^{i+1,j-1})) \cup \mathcal{U}(\mathcal{F}(L^{i-1,j})), \quad (2)$$

where $\mathcal{U}(\cdot)$ denotes the up-sampling layer, and $[\cdot, \cdot]$ denotes the concatenation layer.

The CSAM consists of two cascaded attention units. The feature maps $L^{i,j}$ from node $L^{i,j}$ ($i \in \{0, 1, 2, \ldots, I\}$, $j \in \{0, 1, 2, \ldots, J\}$) are sequentially processed by a 1D channel attention process can be computed as:

$$L^{i,j} = \mathcal{P}_{\text{max}}(\mathcal{F}(L^{i,j})), \quad (1)$$

where $\mathcal{F}(\cdot)$ denotes multiple cascaded convolution layers of the same convolution block. $\mathcal{P}_{\text{max}}(\cdot)$ denotes max-pooling with a stride of 2. When $j > 0$, each node receives outputs from three directions including dense plain skip connection and nested bi-direction interactive skip connection, the stack of feature maps represented by $L^{i,j}$ is generated as:

$$L^{i,j} = \mathcal{F}(L^{i,j}) = \mathcal{P}_{\text{max}}(\mathcal{F}(L^{i+1,j-1})) \cup \mathcal{U}(\mathcal{F}(L^{i-1,j})), \quad (2)$$

where $\mathcal{U}(\cdot)$ denotes the up-sampling layer, and $[\cdot, \cdot]$ denotes the concatenation layer.
Before multiplication, the attention maps $M_i(L)$ are stretched to the size of $M_i(L) \in \mathbb{R}^{C_i \times H_i \times W_j}$.

Similar to channel attention process, the spatial attention process can be summarized as follows:

$$M_i(L') = \sigma \left( f^{3 \times 7}(P_{\text{max}}(L')) \cdot (P_{\text{avg}}(L')) \right),$$

$$L'' = M_i(L') \otimes L',$$

where $f^{3 \times 7}$ represents a convolutional operation with the filter size of $7 \times 7$. The attention maps $M_i(L)$ are also stretched to the size of $M_i(L) \in \mathbb{R}^{C_i \times H_i \times W_j}$ before multiplication.

C. The Feature Pyramid Fusion Module

After the feature extraction module, we develop a feature pyramid fusion module to aggregate the resultant multi-layer features. As shown in Fig. 3 (b), we first upscale multi-layer features to the same size of $L$ features. As shown in Fig. 3 (b), we first upscale multi-pyramid fusion module to aggregate the resultant multi-layer features. After the feature pyramid fusion module, we introduce a robust feature maps to generate global robust feature maps:

$$G = \{L_{\text{en}_\text{up}}^{0,J}, L_{\text{en}_\text{up}}^{1,J}, \ldots, L_{\text{en}_\text{up}}^{L,J} \},$$

D. The Eight-Connected Neighborhood Clustering Module

After the feature pyramid fusion module, we introduce an eight-connected neighborhood clustering module [34] to cluster the pixels belonging to the same target and calculate the centroid of each target. If any two pixels $(m_0, n_0), (m_1, n_1)$ in feature maps $G$ have intersection areas in their eight neighborhoods, i.e.,

$$N_8(m_0, n_0) \cap N_8(m_1, n_1) \neq \emptyset,$$

where $N_8(m_0, n_0)$ and $N_8(m_1, n_1)$ represent the eight neighborhoods of pixel $(m_0, n_0)$ and $(m_1, n_1)$.

### IV. The NUDT-SIRST Dataset

#### A. Motivation

Quality, quantity, and scene diversity of data significantly affect the performance of CNN-based methods. As shown in Table I, existing datasets either lack enough scenes (e.g., NUST-SIRST [32] and CQU-SIRST [7]) or have limited data capacity (e.g., NUAA-SIRST [24]). It is costly to collect a large-scale dataset with accurate pixel-level annotations. These issues hinder the further development of CNN-based methods. Inspired by the solutions in other data-scarcity field (e.g., ship detection [35], [36], moving car detection [37], [38]), we develop a large-scale infrared small target dataset (namely, the NUDT-SIRST dataset). Our NUDT-SIRST dataset contains common multi-target scenarios, more small targets, and less visually salient targets.

#### TABLE I

| Datasets                      | Image Type | Background Scene | #Image | Label Type   | Target Type | Public |
|-------------------------------|------------|------------------|--------|--------------|-------------|--------|
| NUAA-SIRST(ACM) [24]          | real       | Cloud/City/Sea   | 427    | Manual Coarse Label | Point/Spot | ✔️     |
| NUST-SIRST [32]               | synthetic  | Cloud/City/River/Road | 10000 | Manual Coarse Label | Point/Spot | ✔️     |
| CQU-SIRST [7]                 | synthetic  | Cloud/City/Sea   | 1676   | Ground Truth | Point/Spot | ✗      |
| NUDT-SIRST (ours)             | synthetic  | Cloud/City/Sea/Field/Highlight | 1327 | Ground Truth | Point/Spot | ✔️     |

![Comparison of existing public SIRST datasets](image_url)

Fig. 6. Comparison of existing public SIRST datasets on (a) the number of targets, (b) target size, and (c) target brightness. Our NUDT-SIRST dataset contains more multi-target scenarios, more small targets, and less visually salient targets.

The quality, quantity, and scene diversity of data significantly affect the performance of CNN-based methods. As shown in Table I, existing datasets either lack enough scenes (e.g., NUST-SIRST [32] and CQU-SIRST [7]) or have limited data capacity (e.g., NUAA-SIRST [24]). It is costly to collect a large-scale dataset with accurate pixel-level annotations. These issues hinder the further development of CNN-based methods. Inspired by the solutions in other data-scarcity field (e.g., ship detection [35], [36], moving car detection [37], [38]), we develop a large-scale infrared small target dataset (namely, the NUDT-SIRST dataset). Our NUDT-SIRST dataset contains common multi-target scenarios, more small targets, and less visually salient targets.

The quality, quantity, and scene diversity of data significantly affect the performance of CNN-based methods. As shown in Table I, existing datasets either lack enough scenes (e.g., NUST-SIRST [32] and CQU-SIRST [7]) or have limited data capacity (e.g., NUAA-SIRST [24]). It is costly to collect a large-scale dataset with accurate pixel-level annotations. These issues hinder the further development of CNN-based methods. Inspired by the solutions in other data-scarcity field (e.g., ship detection [35], [36], moving car detection [37], [38]), we develop a large-scale infrared small target dataset (namely, the NUDT-SIRST dataset). Our NUDT-SIRST dataset contains common multi-target scenarios, more small targets, and less visually salient targets.

The quality, quantity, and scene diversity of data significantly affect the performance of CNN-based methods. As shown in Table I, existing datasets either lack enough scenes (e.g., NUST-SIRST [32] and CQU-SIRST [7]) or have limited data capacity (e.g., NUAA-SIRST [24]). It is costly to collect a large-scale dataset with accurate pixel-level annotations. These issues hinder the further development of CNN-based methods. Inspired by the solutions in other data-scarcity field (e.g., ship detection [35], [36], moving car detection [37], [38]), we develop a large-scale infrared small target dataset (namely, the NUDT-SIRST dataset). Our NUDT-SIRST dataset contains common multi-target scenarios, more small targets, and less visually salient targets.
Fig. 7. Synthesis process of our dataset. (a) Adaptive target size function. Pre-collected background images are fed into a scene-aware CNN model to identify the type of background. Then, the size and type of candidate targets are selected with pre-defined possibility $P_{size}$. The background and selected targets are directly added. (b) Adaptive intensity and blur function. The initially synthesized images are sequentially fed into an adaptive intensity function $F_{intensity}$ and a Gaussian blur function $F_{blur}$ to make the targets’ intensity and boundary realistic, respectively. (c) Samples of our NUDT-SIRST dataset. Our dataset covers multiple real infrared backgrounds, various target types, rich poses, and ground truth labels. ↗, ↙, →, and ⊗ represents different moving directions of targets.

B. Implementation Details

High-quality synthesized images should be both physically reasonable and visually realistic. To render reasonable images, as shown in Fig. 7(a), we first used a Gaussian kernel function and collected target templates (e.g., spot, plane, ship, and UAV) to simulate point, spot, and extended targets, respectively. Then, we adopted an adaptive target size function $F_{size}$ to make sure the size of target and the combination of virtual targets with real infrared background reasonable. In this function, a scene-aware CNN $F_{scene}$ is first used to identify the type of the background. Then, we assigned pre-defined possibility $P_{size}$ to identify the size and type of candidate targets. In this way, we can avoid the unreasonable combination of target and background such as a big plane target with city background and a ship target with sky background.

To generate visually realistic images, as shown in Fig. 7(b), we used an adaptive intensity function $F_{intensity}$ and a Gaussian blur function $F_{blur}$ to adjust the target’s intensity and blur it’s boundary, respectively. In the adaptive intensity function, we adjusted the average gray value of the target to keep the target’s SCR fixed at an empirical value $C$ (i.e., 3, 4, 5, and 6). That is:

$$SCR = \frac{\mu_T - \mu_B}{\sigma_B} = C,$$

where $\mu_B$ and $\sigma_B$ are the average and standard derivation of the background. Then, we imposed a $5 \times 5$ Gaussian blur function with different $\sigma$ (i.e., 0.2, 0.5, 1.0, etc.) on the images to ensure the smoothness of the synthesized images. Finally, we manually removed visually low-quality images.

C. Comparison to Existing Datasets

In this subsection, we compare our NUDT-SIRST dataset to several public SIRST datasets. Following [24], we use three metrics (i.e., the number of targets, target size, and target brightness) to evaluate these datasets. As shown in Fig. 6(a), about 37% of images in the NUDT-SIRST dataset contain no less than 2 targets. This ratio is much higher than the other two datasets. Target size distribution in Fig. 6(b) shows that 27% of targets occupy no more than 0.01% area of the whole image and 96% of targets meet the SPIE’s defination for small targets (i.e., the target should be smaller than 0.15% area of the whole image). Point and small target ratios are much higher than the other two datasets. As shown in Fig. 6(c), there are about 32% of targets locating outside of top 10% of the image brightness value. It demonstrates that the images of our dataset are less visually salient than other datasets. In summary, compared with existing datasets [24], [32], our dataset introduces more challenging scenes (i.e., multiple targets, point target, and dim target scenes).

V. EXPERIMENT

In this section, we first introduce our evaluation metrics and implementation details. Then, we compare our DNA-Net to several state-of-the-art SIRST detection methods. Finally, we present ablation studies to investigate our network.

A. Evaluation Metrics

Pioneering CNN-based works [24], [31], [32] mainly use pixel-level evaluation metrics like $IoU$, precision, and recall values. These metrics mainly focus on the target
In this paper, we adopted a segmentation network as our baseline to generate a pixel-level segmentation map and then used a clustering algorithm to achieve target localization. The U-net paradigm with ResNets [40] was chosen as our segmentation backbone. The number of down-sampling layer $i$ was chosen as 4. Our network was trained using the Soft-IoU loss function and optimized by the Adagrad method [41] with the CosineAnnealingLR scheduler. We initialized the weights and bias of our model using the Xavier method [42]. We set the learning rate, batch size, and epoch size as 0.05, 16, and 1500, respectively. All models were implemented in PyTorch [43] on a computer with an AMD Ryzen 9 3950X @ 2.20 GHz CPU and an Nvidia GeForce 3090 GPU.

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

To demonstrate the superiority of our method, we compare our DNA-Net to several state-of-the-art (SOTA) methods, including traditional methods (Top-Hat [6], Max-Median [9], WSLCM [13], TLCCM [12], IPI [7], NRAM [16], RIPT [8], PSTNN [17], MSLSITP [5]) and CNN-based methods (MDvsFA-cGAN [32], ACM [24], ALCNet [31]) on the NUAA-SIRST and NUDT-SIRST datasets.\(^1\) For fair comparison, we retrained all the CNN-based methods on the same training datasets as our DNA-Net. It is worth noting that we use our implementations for these methods for fair comparison. Most of these open-source CNN-based codes are rewritten by pytorch and released at: https://github.com/YeRen123455/Infrared-Small-Target-Detection.

1) Quantitative Results: For all the compared algorithms, we first obtained their predicts and then performed noise suppression by setting a threshold to remove low-response areas. Specifically, the adaptive threshold ($T_{adaptive}$) was calculated for traditional methods according to:

$$T_{adaptive} = \text{Max} \left[ \text{Max}(G) \times 0.7, 0.5 \times \sigma(G) + \text{avg}(G) \right]$$

where $\text{Max}(G)$ represents the largest value of output. $T_{adaptive}$ represents adaptive threshold. $\sigma(G)$ and $\text{avg}(G)$ mean the standard derivation and average value of output, respectively. For CNN-based methods, we followed their original papers and adopted their fixed thresholds (i.e., 0, 0, 0.5 for ACM [24], ALCNet [31], and MDvsFA-cGAN [32], respectively). We kept all remaining parameters the same as their original papers.

Quantitative results are shown in Table II. The improvements achieved by our DNA-Net over traditional methods are significant. That is because, both NUDT-SIRST and NUAA-SIRST contain challenging images with different SCR, clutter background, target shape, and target size. Our DNA-Net can learn discriminative features robust to scene variations. In contrast, the traditional methods are usually designed for specific

\(^1\)Note that, we follow ACM [24] and ALCNet [31] to not use the NUST-SIRST for comparison in the main body of our manuscript since only about 30% of targets meet the SPIE’s definition of small targets. To achieve a more comprehensive comparison, we have updated the experimental results of NUST-SIRST and released the trained model at our Github repository.
The best results are in **red** and the second best results are in **blue**. **TR** = 50% means 50% images are used for training and the rest are used for test.

| Method Description          | NUDT-SIRST (TR=50%) | NUAA-SIRST (TR=50%) |
|-----------------------------|---------------------|---------------------|
|                             | IoU ($\times 10^2$) | $P_d$ ($\times 10^2$) | $F_a$ ($\times 10^6$) | IoU ($\times 10^2$) | $P_d$ ($\times 10^2$) | $F_a$ ($\times 10^6$) |
| Filtering Based: Top-Hat [6] | 20.72               | 78.41               | 166.7               | 7.143               | 79.84               | 1012               |
| Filtering Based: Max-Median [9] | 4.197               | 58.41               | 36.89               | 4.172               | 69.20               | 55.33               |
| Local Contrast Based: WSLCM [13] | 2.283               | 56.82               | 1309                | 1.158               | 77.95               | 5446               |
| Local Contrast Based: TLLCM [12] | 2.176               | 62.01               | 1608                | 1.029               | 79.09               | 5899               |
| Low Rank Based: IPI [7]    | 17.76               | 74.49               | 41.23               | 25.67               | 85.55               | 11.47               |
| Low Rank Based: NRAM [16]  | 6.927               | 56.20               | 19.27               | 12.16               | 74.52               | 13.85               |
| Low Rank Based: RPT [5]    | 29.44               | 91.85               | 344.3               | 11.05               | 79.08               | 22.61               |
| Low Rank Based: PSTNN [17] | 14.85               | 66.13               | 44.17               | 22.40               | 77.95               | 29.11               |
| Low Rank Based: MSLSTIPT [5] | 8.342               | 47.40               | 888.1               | 10.30               | 82.13               | 1131               |
| CNN Based: MDvsFA-cGAN [32] | 75.14               | 90.47               | 25.34               | 60.30               | 89.35               | 56.35               |
| CNN Based: ACM [24]            | 67.08               | 95.97               | 10.18               | 70.33               | 93.91               | 3.728               |
| CNN Based: ALCNet [31]       | 81.40               | 96.51               | 9.261               | 73.33               | 96.57               | 30.47               |
| DNA-Net-VGGG10 (ours)       | 85.23               | 96.95               | 6.782               | 74.96               | 97.34               | 26.73               |
| DNA-Net-ResNet10 (ours)     | 86.36               | 97.39               | 6.897               | 76.24               | 97.71               | 12.80               |
| DNA-Net-ResNet18 (ours)     | 87.09               | 98.73               | 4.223               | 77.47               | 98.48               | 2.353               |
| DNA-Net-ResNet34 (ours)     | 86.87               | 97.98               | 3.710               | 77.54               | 98.10               | 2.510               |

**Table II**

| Evaluation Metrics | NUDT-SIRST (TR=50%) | NUAA-SIRST (TR=50%) |
|--------------------|---------------------|---------------------|
| IoU ($\times 10^2$) | 20.72               | 78.41               |
| $P_d$ ($\times 10^2$) | 78.41               | 166.7               |
| $F_a$ ($\times 10^6$) | 166.7               | 7.143               |
| $P_d$ ($\times 10^2$) | 79.84               | 1012               |
| $F_a$ ($\times 10^6$) | 1012               | 7.143               |

2) Qualitative Results: Qualitative results on two datasets (i.e., NUDT-SIRST, NUAA-SIRST) are shown in Fig. 8 and Fig. 9. Compared with traditional methods, our method can produce output with precise target localization and shape segmentation under very low false alarm rate. Nonetheless, the traditional methods only perform well on point targets (e.g., image-3), and easily generate lots of false alarm areas in local highlight areas (e.g., image-4 and image-6). Moreover, as shown in Fig. 12, we divided our NUDT-SIRST dataset into point targets subset, spot targets subset, and extended targets subset. With the increase of spot and extended targets ratio, traditional methods suffer dramatic performance decrease while our DNA-Net maintains high accuracy. That is because, the performance of traditional methods rely heavily on handcrafted features and cannot adapt to the variations of target sizes.

The CNN-based methods (i.e., MDvsFA-cGAN, ACM, and ALCNet) perform much better than traditional methods. However, due to the complicated scenes in our NUDT-SIRST, fusion module help to learn discriminative features to achieve better performance.

Quantitative results in Table IV demonstrate that our method is superior to other deep-learning based methods under different pre-defined deviation thresholds.

**Table III**

Comparison to SOTA methods in terms of **Train Time**, **Inference Time**, and **IoU ($\times 10^2$)**, **$P_d$ ($\times 10^2$)**, **$F_a$ ($\times 10^6$)** on the NUDT-SIRST dataset.

| Method | Evaluation Metrics | Train Time ($\times 10^2$) | Inference Time ($\times 10^2$) | IoU ($\times 10^2$) | $P_d$ ($\times 10^2$) | $F_a$ ($\times 10^6$) |
|--------|---------------------|-----------------------------|-----------------------------|---------------------|---------------------|---------------------|
| MDvsFA-cGAN [32] | 9.952h            | 0.019s                     | 75.14/90.47/25.34        | 75.14/90.47/25.34        | 75.14/90.47/25.34        |
| ACM (ResNet20) [24] | 0.946h            | 0.011s                     | 67.08/95.37/10.18        | 67.08/95.37/10.18        | 67.08/95.37/10.18        |
| ALCNet (ResNet20) [31] | 7.623h            | 0.021s                     | 81.40/96.51/9.261        | 81.40/96.51/9.261        | 81.40/96.51/9.261        |
| DNA-Net-ResNet18-Light | 3.862h            | 0.012s                     | 83.68/97.28/13.23        | 83.68/97.28/13.23        | 83.68/97.28/13.23        |

**Table IV**

$P_d$ ($\times 10^2$) and $F_a$ ($\times 10^6$) values achieved by different state-of-the-art methods on the NUDT-SIRST dataset with different settings of $D_{thresh}$.

| Method | Maximum Centroid Deviation | $D_{thresh}<3$ | $D_{thresh}<3$ | $D_{thresh}<4$ |
|--------|----------------------------|----------------|----------------|----------------|
| MDvsFA-cGAN [32] | 89.31/90.47/25.34 | 90.47/25.34 | 91.21/24.98 |
| ACM (ResNet20) [24] | 95.56/15.65 | 95.97/10.18 | 95.97/10.18 |
| DNA-Net-ResNet18 | 98.33/13.23 | 98.73/4.228 | 98.73/4.228 |
Fig. 8. Qualitative results achieved by different SIRST detection methods. For better visualization, the target area is enlarged in the right-top corner. The correctly detected target, false alarm, and miss detection areas are highlighted by red, yellow, and green dotted circles, respectively. Our DNA-Net can generate output with precise target localization and shape segmentation under a lower false alarm rate.

MDvsFA-cGAN produces many false alarm and miss detection areas (Fig. 9). Our DNA-Net is more robust to these scene changes. Moreover, our DNA-Net can generate better shape segmentation than ALCNet. That is because, our designed new backbone can well adapt to various clutter background, target shape, and target size challenges and thus achieves better performance.

3) Computational Efficiency: In this part, we reduced half of the channels in DNA-Net-ResNet10 to build DNA-Net-ResNet10-Light and compared it to several competitive methods (i.e., MDvsFA-cGAN [32], ACM [24], ALCNet [31]) in terms of training time and inference time. As shown in Table III, our DNA-Net-ResNet10-Light achieves the highest IoU, $P_d$, and the lowest $F_a$ with comparable training and inference time. This clearly demonstrates the high computational efficiency of our method.

D. Ablation Study

In this subsection, we compare our DNA-Net with several variants to investigate the potential benefits introduced by our network modules and design choice.

1) The Dense Nested Interactive Module (DNIM): The dense nested interactive skip-connection module is used to interact with features at different scale levels to enlarge receptive fields while maintain fine-grained features at the finest scale level. To demonstrate the effectiveness of our DNIM, we introduced three network variants and made their model sizes comparable for fair comparison.

Table V shows the comparative results achieved by DNA-Net and its variants. It can be observed that the $IoU$, $P_d$, and $F_a$ values of DNA-Net w/o DNIM suffer decreases of 2.08%, 2.23%, and an increase of $4.298 \times 10^{-6}$ on the NUDT-SIRST dataset. Similar results are also observed on the NUAA-SIRST dataset.

![Table of Image](image_url)
dataset. That is because, DNIM progressively aggregates features at multiple scales to maintain the target information at the finest scale for better performance. Visualization maps shown in Fig. 10 also demonstrates the effectiveness of our DNIM. Small targets are lost in the feature maps of the deep layer in DNA-Net w/o DNIM (i.e., L(4,0), L(3,1)).

- **DNA-Net w/o DNIM**: We replaced the dense nested interactive skip connection module with a regular plain skip connection module.
- **DNA-Net-left-to-right**: As shown in Fig. 11(c), multiple U-shape subnetworks with different depths are stacked from left to right. Each node in the middle part of the network can receive features from its own and the lower layer.
- **DNA-Net-top-to-bottom**: We stacked the U-shape subnetworks from top to bottom to generate DNA-Net-top-to-bottom, as shown in Fig. 11(b). Different from DNA-Net-left-to-right, this variant stacks U-shape subnetworks with three kinds of depth and only its core part uses tri-direction skip connection.

As shown in Table V, DNA-Net-left-to-right suffers decreases of 1.20%, 1.44%, and an increase of $0.426 \times 10^{-6}$ in terms of $I_oU$, $P_d$, and $F_a$ values over DNA-Net on the NUDT-SIRST dataset. That is because, each node in DNA-Net-left-to-right only interacts with the deep layer instead of full interaction among shallow, their-own, and deep layers. Shallow layer has rich localization and profile information, but the information is not fully incorporated at the skip connection stage. Consequently, this variant has limited performance.

As compared to our DNA-Net, the variant DNA-Net-top-to-bottom suffers decreases of 1.34%, 1.77%, and an increase of $3.459 \times 10^{-6}$ in terms of $I_oU$, $P_d$, and $F_a$ values on NUDT-SIRST dataset. That is because, only the core part of this
variant adopts tri-direction skip connection, the remaining part still uses the plain skip connection. Moreover, its tri-direction interactive area is relatively shallow, high-level information cannot be fully exploited at shallow layers.

2) The Channel and Spatial Attention Module (CSAM): The channel and spatial attention module is used for adaptive feature enhancement to achieve better feature fusion. To investigate the benefits introduced by this module, we compare our DNA-Net with four variants. To achieve fair comparison (i.e., comparable model size), we increased the number of filters of all convolution layers of four variants to make their model sizes slightly larger than DNA-Net.

- DNA-Net w/o CSAM: We removed the channel and spatial attention module in this variant and directly concatenate multi-layer features for subsequent process.
- DNA-Net w/o CSAM (Element-wise summation): We replaced CSAM with common element-wise summation in this variant to explore the effectiveness of CSAM. Specifically, we used $1 \times 1$ convolution operation and up-sampling/down-sampling to make features from different layer identical. Then, an element-wise summation is used to achieve multi-layer feature fusion.
- DNA-Net w/o channel attention: We removed the channel attention operation in this variant to evaluate its contribution.
- DNA-Net w/o spatial attention: We canceled the spatial attention operation in this variant to investigate the benefit introduced by spatial attention.

If CSAM is removed, the performance suffers decreases of 1.19%/1.84%, 2.11%/2.11%, and an increase of $1.515 \times 10^{-6}$ in terms of $IoU$, $P_d$, and $F_a$ for DNA-Net w/o CSAM and DNA-Net w/o CSAM $\oplus$ on the NUDT-SIRST and NUAA-SIRST datasets, respectively. Similar results are achieved on the NUAA-SIRST dataset. This clearly demonstrates the importance of the channel and spatial attention module. As shown in Fig. 13, with the help of CSAM, the feature maps from the deep layer of DNA-Net have high response to informative cues and finally results in precise shape segmentation.

As shown in Table VI, DNA-Net w/o channel attention suffers decreases of 0.82%, 1.77%, and an increase of $0.658 \times 10^{-6}$ in terms of $IoU$, $P_d$, and $F_a$ values over DNA-Net on NUDT-SIRST dataset. That is because, channel attention unit in our DNA-Net can better exploit informative channels to enhance the representation capability of features.

If the spatial attention unit is removed, the performance suffers decreases of 0.95%, 2.00%, and an increase of $0.095 \times 10^{-6}$ in terms of $IoU$, $P_d$, and $F_a$ values for DNA-Net on NUDT-SIRST dataset. That is because, infrared small targets are easily immersed in heavy cloud and noise, it is hard to distinguish these small and dim targets from the background. Spatial attention facilitates the network to pay attention to local informative areas and thus produces better results.

3) The Feature Pyramid Fusion Module (FPFM): The feature pyramid fusion module is used to fuse shallow-layer feature with rich spatial information and deep-layer feature with rich semantic information. To investigate the benefits introduced by this module, we compare our DNA-Net with

| Model             | #Params(M) | NUDT-SIRST | NUAA-SIRST |
|-------------------|------------|------------|------------|
| DNA-Net w/o CSAM  | 4.70       | 85.90/96.736/738 | 75.81/96.19/22.12 |
| DNA-Net w/o CSAM $\oplus$ | 4.71 | 85.25/96.626/710 | 75.35/95.82/34.97 |
| DNA-Net w/o CA    | 4.73       | 86.27/96.964/881 | 76.20/96.96/12.69 |
| DNA-Net w/o SA    | 4.73       | 86.14/96.734/128 | 76.69/97.34/10.96 |
| DNA-Net-ResNet18  | 4.70       | 87.09/98.734/223 | 77.47/98.46/335 |

Fig. 10. Visualization map of DNA-Net (row 1) and DNA-Net w/o DNIM (row 2). The feature maps from the deep layer of DNA-Net w/o DNIM loses representation of small targets. It finally results in low values and miss detection in the output layer.

Fig. 11. Three variants of DNIM. (a) DNA-Net w/o DNIM. (b) DNA-Net-top-to-bottom. (c) DNA-Net-left-to-right. (d) DNA-Net, each color represents different U-shape sub-networks.

Table VI $IoU(\times 10^2)/P_d(\times 10^2)/F_a(\times 10^6)$ VALUES ACHIEVED BY MAIN VARIANTS OF DNA-NEt AND CSAM ON THE NUDT-SIRST AND NUAA-SIRST DATASETS. $\oplus$ MEANS ELEMENT-WISE SUMMING AS FEATURE FUSION METHOD.
three variants, we increased the number of filters of all convolution layers of three variants to make their model sizes comparable for fair comparison.

- **DNA-Net w/o FPFM**: We replaced the feature pyramid fusion module in this variant and only used the output from the final layer as final result.
- **DNA-Net w/o L345**: We removed the outputs of layer 3, 4, and 5 from FPFM in this variant to evaluate the contribution of features from middle and deep layers.
- **DNA-Net w/o L45**: We removed the outputs of layer 4, and 5 from FPFM in this variant to investigate the contribution of features from deep layers.
- **DNA-Net w/o L5**: We removed the outputs of layer 5 from FPFM in this variant to investigate the benefit introduced by the deepest layer of the network.

As shown in Table VII, DNA-Net w/o FPFM suffers decreases of 0.83%, 1.89%, and a increase of $3.01 \times 10^{-6}$ in terms of $IoU$, $P_d$, and $F_a$ on the NUDT-SIRST dataset. Similar results can also be observed on the NUAA-SIRST dataset. That is because, FPFM helps to achieve multi-layer features fusion. The representation from shallow layers and deep layers can be both extracted and fused to generate more robust feature maps as output.

When we gradually removed partial outputs of FPFM from bottom to the top layer, our network suffers decreases of 0.23%, 0.84%, and an increase of $3.01 \times 10^{-6}$ in terms of $IoU$, $P_d$, and $F_a$ for DNA-Net w/o L5. Similar results can also be observed on DNA-Net w/o L45 and DNA-Net w/o L345. That is because, NUDT-SIRST contains rich multi-target scenarios, more small size targets, and less visually salient targets.

Our network can fully fuse low-level and high-level information and thus achieves better performance on NUDT-SIRST.

### E. Benefits of the Synthesized Dataset

In this section, we evaluate the benefits of our synthesized dataset for real IRST tasks. Specifically, we mixed real SIRST images (from the training set of NUAA-SIRST) and synthesized SIRST images (from the training set of NUDT-SIRST) with different ratios to train the networks and evaluated their performance on the real images (from the test set of NUAA-SIRST). As shown in Table VIII, with small ratio of real images, both DNA-Net and ACM can achieve comparable results to baseline results (trained on all real images). That is because, our synthesized dataset can well cover the main challenges for infrared small target detection (i.e., different SCR, clutter background, target shape, and target size). Consequently, the huge cost for collecting real SIRST images can be reduced.
TABLE VIII

| #Real image (x10^5) | #Synthesized image (x10^5) | Method | DNA-Net | ACM |
|---------------------|-----------------------------|--------|---------|-----|
| 0% (0/791)          | 100% (791/791)              |        | 66.84/95.43/28.59 | 60.43/91.25/20.75 |
| 5.3% (42/791)       | 94.7% (749/794)             |        | 70.44/91.19/15.40 | 63.94/92.87/27.37 |
| 10.7% (85/791)      | 89.3% (706/794)             |        | 74.58/97.45/26.23 | 66.69/93.01/16.95 |
| 16.2% (128/791)     | 83.8% (663/794)             |        | 77.23/94.34/40.01 | 69.29/95.06/9.02 |
| 100% (213/213)      | -                           |        | 77.47/94.48/223  | 70.33/93.91/3.72 |

Moreover, we compared the output of our network trained on the mixed dataset with the manually labeled masks of NUAA-SIRST in Fig. 14. It can be observed that the outputs of our network have more reasonable shape segmentation than ground truth labels. That is because, the synthesized SIRST images have absolutely precise labels. The network can learn the essence of infrared small targets with sufficiently well labeled data and finally contribute to the improvement of real SIRST images. Our network can generate better visual performance than ground truth label.

VI. CONCLUSION

In this paper, we propose a DNA-Net to achieve SIRST detection. Different from existing CNN-based SIRST detection methods, we explicitly handle the problem of small targets being lost in deep layers by designing a new tri-direction dense nested interactive module with a cascaded channel and spatial attention model. The intrinsic information of small targets can be incorporated and fully exploited by repeated fusion and enhancement. Moreover, we develop an open SIRST dataset to evaluate the performance of infrared small target detection with respect to challenging scenes. We also reorganized a set of evaluation metrics. Experiments on both our dataset and the public dataset have shown the superiority of our method over the state-of-the-art methods.

REFERENCES

[1] M. Teutsch and W. Kruger, “Classification of small boats in infrared images for maritime surveillance,” in Proc. Int. WaterSide Secur. Conf., Nov. 2010, pp. 1–7.
[2] X. Ying et al., “MoCoPnet: Exploring local motion and contrast priors for infrared small target super-resolution,” 2022, arXiv:2201.01014.
[3] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, “Small infrared target detection based on weighted local difference measure,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 7, pp. 4204–4214, Jul. 2016.
[4] T. Ma, Z. Yang, J. Wang, S. Sun, X. Ren, and U. Ahmad, “Infrared small target detection network with generate label and feature mapping,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
[5] 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.
[6] J.-F. Rivest and R. Fortin, “Detection of dim targets in digital infrared imagery by morphological image processing,” Opt. Eng., vol. 35, no. 7, pp. 1886–1893, Jul. 1996.
[7] C. Q. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, “Infrarot-patch-image model for small target detection in a single image,” IEEE Trans. Image Process., vol. 22, no. 12, pp. 4996–5009, Dec. 2013.
[8] 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.
[9] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, “Max-mean and max-max median filters for detection of small targets,” Proc. SPIE, vol. 3809, pp. 74–83, Oct. 1999.
[10] 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.
[11] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, “A robust infrared small target detection algorithm based on human visual system,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 12, pp. 2168–2172, Dec. 2014.
[12] J. Han, S. Morandi, 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.
[13] 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.
[14] S. Kim and J. Lee, “Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track,” Pattern Recognit., vol. 45, no. 1, pp. 393–406, Jan. 2012.
[15] X. Wang, G. Lv, and L. Xu, “Infrared dim target detection based on visual attention,” Infr. Phys. Technol., vol. 55, no. 6, pp. 513–521, Nov. 2012.
[16] L. Zhang, L. Peng, T. Zhang, S. Cao, and Z. Peng, “Infrared small target detection via non-convex rank approximation minimization joint l2,1 norm,” Remote Sens., vol. 10, no. 11, p. 1821, 2018.
[17] 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.
[18] H. Zhu, S. Liu, L. Deng, Y. Li, and F. Xiao, “Infrared small target detection via low-rank tensor completion with top-hat regularization,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 2, pp. 1004–1016, Oct. 2020.
[19] Y. Dai, Y. Wu, Y. Song, and J. Guo, “Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values,” Infr. Phys. Technol., vol. 81, pp. 182–194, Mar. 2017.
[20] M. Liu, H.-Y. Du, Y.-J. Zhao, L.-Q. Dong, and M. Hui, “Image small target detection based on deep learning with SNR controlled sample generation,” in Current Trends in Computer Science and Mechanical Engineering, vol. 1, Poland, U.K.: De Gruyter Open, 2018, pp. 211–220.
[21] B. McIntosh, S. Venkataramanan, and A. Mahalanobis, “Infrared target detection in cluttered environments by maximization of a target to clutter ratio (TCR) metric using a convolutional neural network,” IEEE Trans. Aerosp. Electron. Syst., vol. 57, no. 5, pp. 3737–3767, Aug. 2021.
[22] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” 2015, arXiv:1506.01497.
[23] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767.
[24] 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), Jan. 2021, pp. 950–959.
[25] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.
[26] J. Zhang, Y. Jin, J. Xu, X. Xu, and Y. Zhang, “MDU-Net: Multi-scale densely connected U-net for biomedical image segmentation,” 2018, arXiv:1812.00352.
Y. Dai, Y. Wu, F. Zhou, and K. Barnard, “Attentional local contrast
K. Wu, E. Otoo, and A. Shoshani, “Optimizing connected component
W. Zhang, M. Cong, and L. Wang, “Algorithms for optical weak small
H. Wang, L. Zhou, and L. Wang, “Miss detection vs. false alarm: Adver-
S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional
et al.
H. Huang et al.
IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 32, 2023
K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for
Q. Yin
C. Xiao
A. Paszke
PYTORCH: An imperative style, high-performance deep

Longguang Wang received the B.E. degree in electrical
engineering from Shandong University (SDU), Jinan, China, in 2015, and the M.E. degree in
information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2017, where he is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology. His research interests include low-level vision and deep learning.

Yingqian Wang received the B.E. degree in electrical
engineering from Shandong University (SDU), Jinan, China, in 2016, and the M.E. degree in
information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018, where he is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology. His research interests include low-level vision, particularly on light field imaging, and image super-resolution.

Zhaiqing Lin received the B.Eng. and Ph.D. degrees
from the National University of Defense Technology (NUDT) in 2007 and 2012, respectively. He is currently an Associate Professor with the College of Electronic Science and Technology, NUDT. His current research interests include infrared image processing and signal processing.

Miao Li received the M.E. and Ph.D. degrees
from the National University of Defense Technology (NUDT) in 2012 and 2017, respectively. He is currently an Associate Professor with the College of Electronic Science and Technology, NUDT. His current research interests include infrared dim and small target detection.

Wei An received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 1999. She was a Senior Visiting Scholar at the University of Southampton, Southampton, U.K., in 2016. She is currently a Professor with the College of Electronic Science and Technology, NUDT. She has authored or coauthored over 100 journals and conference publications. Her current research interests include signal processing and image processing.

Yulan Guo (Senior Member, IEEE) received the B.E.
and Ph.D. degrees from the National University of Defense Technology (NUDT) in 2008 and 2015, respectively. He has authored over 100 papers at highly referred journals and conferences. His current research interests include 3D vision, particularly on 3D feature learning, 3D modeling, 3D object recognition, and scene understanding. He was an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IET Computer Vision, IET Image Processing, and Computers & Graphics. He also served as an Area Chair for CVPR 2021, ICCV 2021, and ACM Multimedia 2021. He organized several tutorials, workshops, and challenges in prestigious conferences, such as CVPR 2016, CVPR 2019, ICCV 2021, 3DV 2021, CVPR 2022, ICFR 2022, and ECCV 2022. He is a Senior Member of ACM.

Boyang Li received the B.E. degree in mechanical
design manufacture and automation from Tianjin
University, China, in 2017, and the M.S. degree in
biomedical engineering from the National Innovation
Institute of Defense Technology, Academy of
Military Sciences, Beijing, China, in 2020. He is cur-
tently pursuing the Ph.D. degree in information and
communication engineering with the National Uni-
versity of Defense Technology (NUDT), Changsha,
China. His research interests include infrared small
target detection, weakly supervised semantic
segmentation, and deep learning.

Chao Xiao received the B.E. degree in commu-
nication engineering and the M.E. degree in
information and communication engineering from the
National University of Defense Technology (NUDT), Changsha, China, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree with the College of Electronic Science.
His research interests include deep learning, small
object detection, and multiple object tracking.