Moderately Dense Adaptive Feature Fusion Network for Infrared Small Target Detection

Chengyu Li, Yan Zhang, Zhiguang Shi, Yu Zhang, and Yi Zhang

***Abstract*— Detecting infrared small targets quickly and accurately in complex backgrounds has always been a challenging task. Data-driven methods have achieved good results because of their powerful feature extraction capabilities. Many algorithms use ResNet or VGG as their backbone, but because of the small size and inconspicuous features, pooling layers in their networks could lead to the loss of targets in deep layers. Even though dense network structure is proposed to alleviate this issue, its excessive dense connections makes it difficult to achieve real-time detection. To meet the requirements of both accurate performance and real-time detection, we propose moderately dense adaptive feature fusion network (MDAFNet). We design a moderately dense adaptive feature fusion (MDAF) module that contains only three feature layers as the backbone of the network. This module connects all the internal features with each other and uses a weighted sum of different layers as the output, promoting feature reuse and maintaining infrared small target features in the deep layers of the network. We also design a coarse-to-fine detection head (CFHead) and introduce auxiliary loss to enable the network to predict target contours with greater precision. Moreover, we propose a new data augmentation method that effectively enhances the generalization performance of network. Experimental results demonstrate that our network achieves excellent performance in detection accuracy and meets the requirements for real-time detection on RTX3080 GPU.**

***Index Terms*— Coarse-to-fine detection head (CFHead), data augmentation, infrared small target detection, moderately dense adaptive feature fusion (MDAF) module, real-time detection.**

I. Introduction

INFRARED detection technology is unique in its ability to detect thermal radiation and operates without interference from lighting [1]. It has a wide range of applications, .including seekers in interceptors [2], tracking systems [2], security nighttime monitoring [3], and aerospace imaging [1]. However, the detection of infrared small targets faces several challenges. First, IR radiation significantly attenuates over distance which causes targets to appear very dim and easily be overwhelmed by background noise [1]. Second, IR small targets have limited texture and structural information [4]. Third, factors such as climate conditions and atmospheric turbulence negatively impact the quality and clarity of infrared images, further increasing the difficulty in small target detection [5].

To detect infrared small targets, many traditional methods have been proposed, including local-contrast-based methods [6], [7], [8], [9], which are primarily applicable for detecting targets that exhibit clear contrast against the background, but may lead to a high number of false alarms due to the presence of high-contrast noise points, human-visual-based methods [10], [11], [12], [13], which are suitable for single-scene object detection but are difficult to achieve target detection in complex backgrounds, and low-rank-based methods [14], [15], [16], [17], [18], which fit in low signal-to-clutter ratio (SCR) but require prior establishment of a model that adopts to the background features and is sensitive to initial conditions, demanding expert knowledge and a lot of engineering efforts. In summary, the reliance on handcrafted features in traditional algorithms has limited their application to specific detection scenarios, thus leading to poor generalization performance.

Different from traditional algorithms, convolutional neural network (CNN)-based methods can automatically learn feature representations and exhibit superior generalization performance [19], [20]. For target detection, there are two types of detection methods, one is based on rotation or horizontal bounding box detection, such as SDANet [21] and MidNet [22]. The other is based on segmentation detection. For example, fully convolutional network (FCN), proposed by Long et al. [23], replaces the traditional fully connected layers with fully convolutional layers and has achieved pixel-level detection by incorporating the upsampling operation into the network. Ronneberger et al. [24] develop a novel architecture called UNet, which preserves more spatial information and fine-grained details by introducing skip connections between the encoding and decoding stages. While the current CNN-based methods have shown good results, there are still three issues remaining. First, the lack of structural features in infrared small targets means that the existing data augmentation techniques, such as flipping, cropping, and rotating, are challenging to effectively enhance the diversity of the data and insufficient for helping network improve generalization performance. Second, pooling layers in the existing networks algorithms could lead to the loss of targets, which makes semantic information of infrared small targets often insufficiently expressed or overwhelmed in deep layers. Although Li et al. [25] propose a dense nested network structure called DNANet to alleviate this issue, excessive dense connection makes it difficult to meet real-time detection requirement.

Manuscript received 14 December 2023; revised 20 February 2024; accepted 18 March 2024. Date of publication 25 March 2024; date of current version 5 April 2024. This work was supported by the National Natural Science Foundation of China under Grant 61302145. *(Corresponding author: Yan Zhang.)*

The authors are with the National Key Laboratory of Science and Technology on Automatic Target Recognition, College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: 1912881096@qq.com; atrthreefire@sina.com; szgstone75@sina.com; zhangyu17d@nudt.edu.cn; zhangyizy@nudt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3381006

Third, the edge and shape information of infrared targets is critical but the existing methods only detect the presence of small targets in infrared images with blurred target contours.

In this article, a CNN-based method moderately dense adaptive feature fusion network (MDAFNet) is proposed with three key components. First, to address the issue that the existing data augmentation methods are not effective for infrared small target detection, we propose multicontrast under different brightness data augmentation method. This data augmentation method expands one image into two brightness levels. In both dark and bright scenes, the contrast of the images is divided into six levels (0–100, 0–150, 0–200, 150–255, 100–255, and 50–255), and a global contrast stretching is applied (0–255), increasing the data number by eight times. Second, we design a moderately dense adaptive feature fusion (MDAF) module as the backbone of our network. The MDAF module maximizes the information flow between layers by directly connecting layers with all the preceding feature maps, and this dense connection strengthens feature propagation and enhances feature reuse, thus alleviating the insufficient expression or overwhelming of infrared small targets in deep layers. What is more, unlike the basic block in ResNet [26], MDAF module never combines features through summation before they are passed into a layer; instead, we use 1×1 convolution to assign weights to different layers for weighted summation, and the values of 1×1 convolution can be learned, which enables the network pay more attention to important features. Third, we design a new detection head called the coarse-to-fine head (CFHead) to obtain the segmentation result. In comparison to the traditional FCNHead, we initially generate an initial coarse prediction and introduce auxiliary loss on it to make it closer to the ground truth; subsequently, we use dilated convolutions with different rates to further refine the coarse prediction and obtain the final result.

The contributions of this article can be summarized as given below.

1) To effectively enhance the network's generalization performance, we develop multicontrast under different brightness data augmentation method that increases the data number by eight times.
2) To alleviate the insufficient expression of infrared small targets in deep layers, we design a moderately dense adaptive fusion module as the backbone of our network, and the dense connection within the MDAF module strengthens feature propagation and enhances feature reuse.
3) To predict target contours with great precision, we design a CFHead. The auxiliary loss is used to help the network generate a coarse initial results, and dilated convolutions with different rates are used to refine coarse prediction.

Extensive experiments are conducted to evaluate our proposed method. Compared with other methods, our approach achieves extremely high accuracy while ensuring real-time performance. The rest of this article is organized as follows. In Section II, the related work is reviewed. The details of the method are introduced in Section III. Then, we illustrate and analyze the experimental results in Section IV. Finally, the conclusion is presented in Section V.

II. Relative Work

A. Infrared Small Target Detection

Infrared small target detection has been continuously studied due to its immense practical value and wide range of applications. Before the rise of CNNs, researchers typically used model-driven methods to directly detect targets or indirectly detect targets by detecting backgrounds. Traditional algorithms included local-contrast-based methods [6], [7], [8], [9], human-visual-based methods [10], [11], [12], [13], low-rank-based methods [14], [15], [16], [17], [18], and filtering-based methods [14], [27]. The reliance on handcrafted features and initialization parameters makes them only suitable for target detection in specific backgrounds, leading to poor generalization performance. With the rise of CNN, some CNN-based methods have emerged. Dai et al. [30] pioneered the development of a segmentation-based network by introducing an innovative asymmetric contextual module, which was designed to effectively combine features from both shallow and deep layers. Furthermore, they enhanced the ACM by integrating a dilated local contrast measure and involved the implementation of a feature cyclic shift scheme to enable a trainable local contrast measure. Li et al. [25] alleviated the issue of insufficient expression of small objects in deep features by designing a dense nest network. However, the excessively dense connections among different blocks made it challenging to achieve real-time detection performance. Kou et al. [28] proposed a lightweight network called IRSTNet which was designed to be lightweight enough to enable real-time detection on most deployment platforms but did not achieve high accuracy. In most of these networks, the basic block in ResNet is used to extract features. But the direct residual connection is actually the simplest linear feature layer fusion and the internal features within the block are not fully used, leading to insufficient expression of advanced semantic information of the targets in deep neural networks. To address this issue, we design an MDAF module.

What is more, in the current segmentation networks for infrared small target detection, most of them use FCNHead as the detection head, which simply upsamples the feature map to the size of the original image to fulfill the requirement of pixelwise segmentation. To obtain better segmentation contour results, we have designed a coarse-to-fine head as the detection head of our network. In addition, we have introduced auxiliary losses in our detection head to further improve the prediction of object contours.

B. Data Augmentation

The data-driven approach has been proven to be far more effective than the model-driven approach in object detection. However, the effectiveness of data-driven methods highly relies on the quantity and quality of data. Unfortunately, in infrared small target detection, acquiring a large volume of data is costly. To address this issue, data augmentation serves

TABLE I
SUMMARY AND INTRODUCTION OF PUBLIC DATASETS

Dataset	Number	Image type	Image size	Description
NUST[29]	10000	synthetic	125×125	The simulated targets are mostly spherical in shape, with significant differences from real-world targets.The backgrounds include clouds, cities, bodies of water, and roads.
NUAA[30]	427	real	-	This is the first publicly available real-world dataset captured under natural conditions, with varying image sizes and backgrounds that include clouds, cities, and oceans.
NUDT[25]	1327	synthetic	256× 256	This is a high-quality simulated dataset that closely resembles real-world scenarios with backgrounds that include clouds, cities, bodies of water, fields, and high altitudes.
IRSTD-1k[5]	1000	real	512× 512	This is a hand-labeled real-world dataset with backgrounds that include clouds, cities, bodies of water, and grassy fields.
IRSTD-1k-ours	1176	real	256× 256	This is a manually refined dataset based on the IRSTD-1k dataset with image sizes of 256x256.

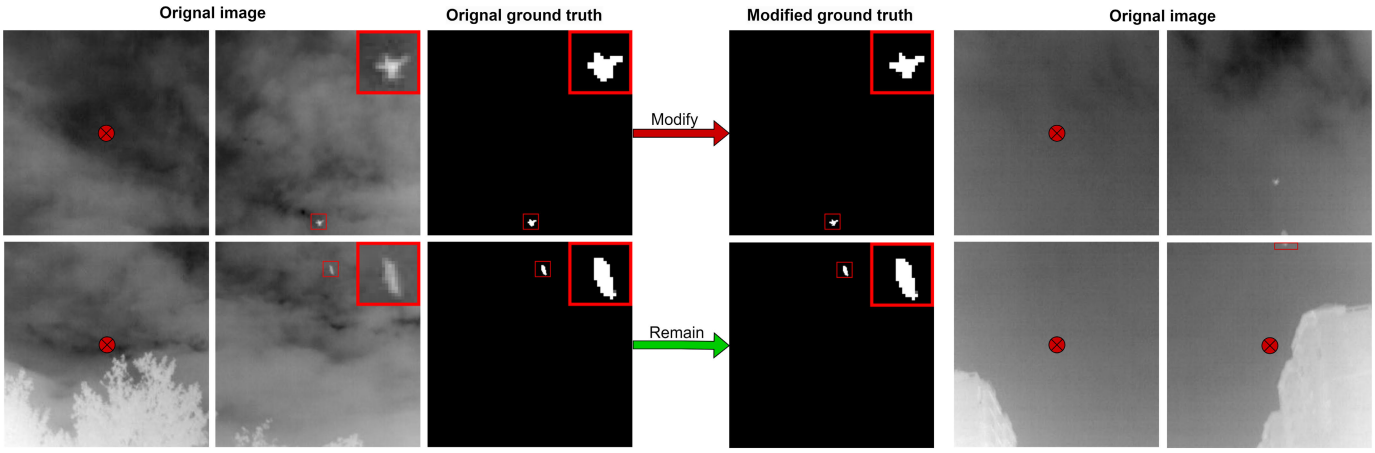

Fig. 1. Illustration of image processing in the IRSTD-1k dataset. Each image is divided into four sections with the center as the reference point. The red cross indicates the deleted portions. To enhance the visibility of the target contours information, the target is enlarged and positioned in the top-right corner of the image.

as an effective solution. Popular data augmentation techniques are as follows.

1) *Traditional Transformations:* Histogram equalization, enhancing contrast or brightness [31], adding the noise to images [32], rotation and image translation, cropping, zooming, random erasing.
2) *Generative Adversarial Networks:* Text-to-image synthesis [33], super-resolution [34] (generating high-resolution image out of low-resolution one), image-to-image translation [35] (e.g., convert sketches into images), image blending [36] (mixing selected parts of two images to get a new one).

Along with data augmentation, the network can achieve more complex representations of the data, thereby improving the generalization performance. But in the case of infrared images, where targets are small and lack texture and structure information, it becomes challenging to effectively diversify the samples using the aforementioned methods. Infrared images captured by infrared detectors exhibit significant differences in contrast and brightness due to variations in detector wavelength, weather conditions, and time periods. Based on this consideration, we propose a new data augmentation method to enable the network to adopt to infrared small target detection with varying brightness and contrast.

C. Datasets

With the development of single-frame detection of infrared small targets, there are currently several datasets available, as shown in Table I. The IRSTD-1k dataset consists of 1000 high-quality infrared images captured in real-world scenarios, the mask labels are manually annotated, but it is found that some of the masks in IRSTD-1k exhibited issues such as blurry target outlines, fragmented labeling regions, and expanded labeling areas beyond the actual target boundaries. The quality of dataset annotations has a significant impact on both training and evaluation accuracy, particularly in infrared small target detection where the proportion of target pixels is small. Even the slightest annotation deviation can potentially result in considerable evaluation bias. For example, a 10-pixel target with 1 pixel annotation error can lead to an approximate 10% deviation. To ensure the reliability of the experimental results, we conduct corrections on the IRSTD-1k dataset to enhance the annotation quality. As shown in Fig. 1, we divide each image into four parts and subsequently remove images that not contain targets or had targets located at the edges. In addition, we reannotate the images with low annotation quality and finally get 1176 high-precision annotated 256 × 256 images. We name the modified dataset IRSTD-1k-ours.

III. METHODOLOGY

A. Overview and Design Strategy

Considering the characteristics of infrared small target, we propose an MDAFNet for infrared small target recognition. The following serve as the design strategies for our network.

1) To achieve real-time detection and prevent the loss of small targets in deep layers, we ensure that the

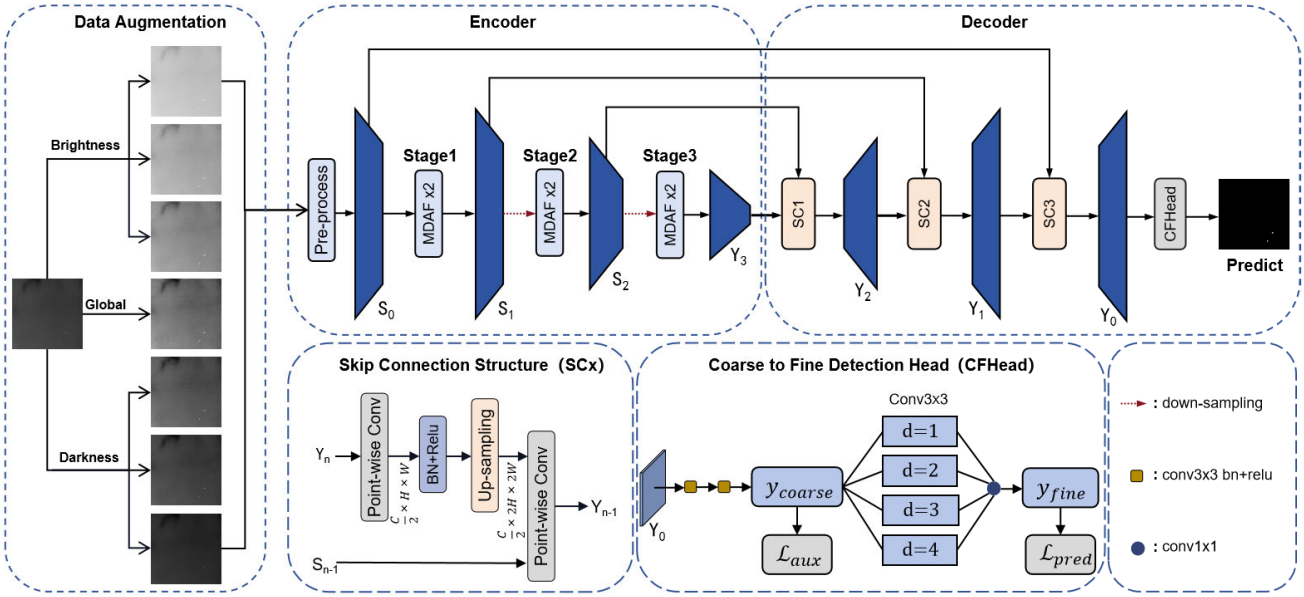

Fig. 2. Illustration of MDAFNet. The encoding of the network is divided into three stages. In each stage, the last layer of the first MDAF module outputs a channel that is twice the number of the input channel, and the other layers keep the same channel as the input channel. Moreover, downsampling is applied between every two stages to reduce the size of feature maps by half, and SCx denotes the 1×1 convolution.

network depth is not excessively deep and the number of downsampling operations is limited.

2) To obtain effective high-level semantic information features and enhance the reuse of low-level information, we design the MDAF module as our network backbone and adopt skip connections with 1×1 convolutions to incorporate contextual information.
3) To enhance the network's segmentation capability for target contours, we replace the traditional FCNHead with CFHead and introduce auxiliary loss.

As illustrated in Fig. 2, our MDAFNet takes a single-frame infrared small target image as the input. It undergoes a sequential process, starting with the encoding phase for extracting image features, followed by the decoding phase for incorporating multilevel feature information. In the encoding phase, features are extracted through three stages, each stage containing two MDAF modules. Downsampling is performed between every two stages, and the number of channels of the feature maps is doubled while the size of the feature maps is halved after each stage. In the decoding phase, the deep feature maps are upsampled and the number of channels is halved, and then incorporated with adjacent shallow information through 1×1 convolution. Finally, the CFHead is used to generate the segmentation results.

Section III-B introduces the proposed MDAF module in the encoding phase. Section III-C introduces the skip connection structure in the decoding phase. Section III-D introduces the proposed coarse to fine feature detection head and auxiliary loss. Section III-E describes the proposed multicontrast under different brightness data augmentation method.

B. Moderately Dense Adaptive Feature Fusion Module

Infrared small targets are easily overwhelmed in high-level features. To address this issue, we design the MDAF module which adopts the dense connection strategy in the basic block of ResNet. Because our network has a depth of only 21, we used larger convolutional kernels in the first layer of the MDAF module to achieve a larger receptive field. As shown in Fig. 2, we initially process the input image using 3×3 convolution with start channels of k, and the output is denoted as S_0, which can be computed as

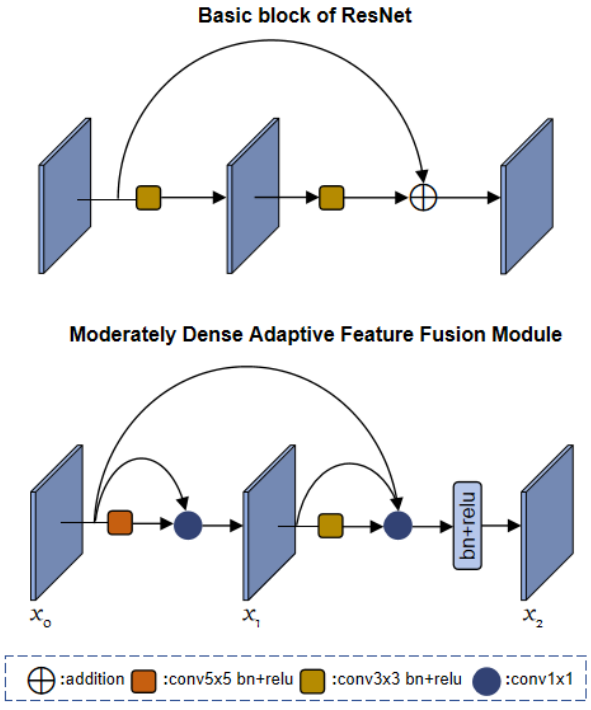

Fig. 3. Illustration of the basic block of ResNet and MDAF.

$$S_0 = F^{3\times3}(\text{input}) \tag{1}$$

where $F^{3\times3}(\cdot)$ denotes the 3×3 convolution followed by batch normalization and relu. Next, different from the basic block in ResNet, we introduce dense connections within the block

to fully use its internal feature maps, as shown in Fig. 3, and the MDAF module consists of three layers: input layer x_0, intermediate layer x_1, and output layer x_2, x_1 can be computed as

$$x_1 = f^{1\times1}\left(\text{concat}\left[x_0, F^{5\times5}(x_0)\right]\right) \tag{2}$$

where concat[·] denotes the concatenation operation, $f^{1\times1}(\cdot)$ denotes the 1×1 convolution, and $F^{5\times5}(\cdot)$ denotes the 5×5 convolution followed by batch normalization and relu. In the output part of MDAF module, instead of using residual connections, we use concatenation followed by 1×1 convolution, this is because a 1×1 convolution acts as an adaptive linear combination of input feature maps, while residual connections are just a specific case of linear combinations. By learning the weights of the 1×1 convolution, the network can adaptively allocate weights to each feature map, thereby increasing the flexibility of incorporating features, and the output layer x_2 can be computed as

$$x_2 = F^{1\times1}\left(\text{concat}\left[x_0, x_1, F^{3\times3}(x_1)\right]\right). \tag{3}$$

C. Skip Connection Structure

After the input image undergoes three stages and produces the respective output feature maps S_1, S_2, and Y_3. To fully use the feature maps generated by the network, in the network encoder phase, we adopt the skip connection strategy from UNet [24], but the difference is that we use only 1×1 convolution kernels for feature integration, enabling adaptive linear combinations of features across all the levels, and the nonlinear factor involved in feature combination is introduced through relu function. As illustrated in Fig. 2, given feature maps Y_n and S_{n-1} as inputs, we use a 1×1 convolutional kernel to reduce the channel of Y_n by half and perform upsampling using the nearest-neighbor interpolation method, doubling the feature maps size. At this point, both processed Y_n and S_{n-1} possess the same channel and size. They are fed into the 1×1 convolution to achieve adaptive linear fusion, resulting in the output Y_{n-1}, which can be computed as

$$Y_{n-1} = f^{1\times1}\left(\text{concat}\left[\text{Up}\left(F^{1\times1}(S_n)\right), Y_{n-1}\right]\right) \tag{4}$$

where $F^{1\times1}(\cdot)$ denotes the 1×1 convolution followed by batch normalization and relu, Up(·) denotes the upsampling operation, and $f^{1\times1}(\cdot)$ denotes the 1×1 convolution. The number of channels in the output Y_{n-1} keeps the same as that in S_{n-1}.

D. CFHead and Auxilary Loss

After the decoding stage, as shown in Fig. 2, we design a CFHead to generate the final segmentation result. To obtain a more precise prediction of the target contours, we first allow the network to generate a coarse prediction result y_{coarse}

$$y_{\text{coarse}} = F_2^{3\times3}\left(F_1^{3\times3}(Y_0)\right). \tag{5}$$

The channel number of $F_1^{3\times3}$ is four times Y_0, and the $F_2^{3\times3}$ is 1. Then, we apply 3×3 convolutions with dilation rates of 1, 2, 3, and 4 to y_{coarse}. The four output feature maps are then summed together to generate the final prediction result denoted as y_{fine}

$$y_{\text{fine}} = \sum_{i=1}^{4} d_i(y_{\text{coarse}}) \tag{6}$$

where $d_i(\cdot)$ denotes the dilated convolution with a dilation rate of i and a kernel size of 3×3.

To handle the imbalance between infrared small target and background, we adopt the Soft-IoU [37] loss function for this highly unbalanced segmentation task. As illustrated in Fig. 2, we let the two loss functions pass through all the previous layers. The auxiliary loss helps optimize the learning process, while the primary branch loss takes the most responsibility. We add weight α to balance the auxiliary loss. The total loss of the network is denoted as $\mathcal{L}_{\text{total}}$

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{aux}}(y_{\text{coarse}}, \text{label}) + \mathcal{L}_{\text{pred}}(y_{\text{fine}}, \text{label}) \tag{7}$$

where $\mathcal{L}(\cdot)$ is the Soft-IoU loss, and label represents the true annotated mask of the image.

E. Multicontrast Under Different Brightness Data Augmentation Method

Infrared images have various contrast and brightness in different scenes. To enrich the diversity of data and improve the generalization performance of the network, we apply different stretching ranges to the original infrared image including low-level brightness, high-level brightness, and global stretching. We linearly interpolate pixel values of original images into several different ranges (0–L_1, …, 0–L_l, 0–255, H_1–255, …, H_h–255), where L_i and H_i, respectively, represent the maximum and minimum values of the interpolation interval for the low-level brightness and high-level brightness. In this article, we have selected seven ranges, which are as follows: (0–100, 0–150, 0–200, 0–255, 200–255, 150–255, and 100–255). The augmented images are denoted as $I_{(i,j)}$

$$I_{(i,j)} = i + (I_0 - \min) \times \frac{j - i}{\max\text{-}\min} \tag{8}$$

where I_0 is the original image, max and min, respectively, represent the maximum and minimum values of the original image pixels, and i and j, respectively, represent the minimum and maximum values of the pixel interpolation range.

IV. Experiment

In this section, we first introduce our evaluation metrics and implementation details. Then, we compare our method with several sirst detection methods. Finally, we present ablation studies to investigate our network.

A. Evaluation Metrics and Implementation Details

We compare the proposed MDAFNet with other methods using several common metrics.

TABLE II

COMPARISON OF DIFFERENT INFRARED SMALL TARGET RECOGNITION ALGORITHMS BASED ON THE METRICS OF nIOU(10^2), P_d(10^2), F_a(10^5). LARGER VALUES OF nIOU AND P_d INDICATE BETTER PERFORMANCE, WHILE SMALLER VALUES OF F_a ARE DESIRABLE. THE ALGORITHMS WITH THE BEST PERFORMANCE AMONG THE COMPARED ALGORITHMS ARE HIGHLIGHTED IN RED FONT. IN THE FINAL GENERATED PREDICTION MAP, A THRESHOLD OF 0.5 IS SET, WHERE PIXEL VALUES GREATER THAN 0.5 ARE CONSIDERED AS TARGETS, AND PIXEL VALUES LESS THAN 0.5 ARE CONSIDERED AS BACKGROUND

Year	Method	NUDT			IRSTD-1k			IRSTD-1k-ours		
		$nIoU$	P_d	F_a	$nIoU$	P_d	F_a	$nIoU$	P_d	F_a
2010	NEW_TOPHAT[27]	25.25	17.89	20.34	56.35	14.90	2.54	3.54	2.28	19.17
2013	IPI[15]	39.54	51.25	6777.99	30.33	43.61	9592.51	4.47	10.54	5.87
2019	PSTNN[16]	19.07	17.79	7.00	27.83	23.32	8.56	2.28	2.59	24.15
2021	FKRW[38]	24.24	17.28	9.41	16.16	10.60	2.21	2.28	1.33	17.00
2021	MDWCM[39]	9.13	6.03	1.882	6.29	3.48	0.40	0.84	0.54	5.03
2015	UNet[24]	88.53	91.70	1.87	55.30	62.13	16.47	69.05	75.24	16.19
2020	ALCNet[40]	82.57	87.29	3.07	56.37	63.43	19.79	73.4	76.65	7.78
2021	LSPM[41]	44.41	57.18	14.42	21.57	28.75	2.62	40.63	56.08	16.91
2021	ACM-FPN[30]	67.17	79.12	16.36	52.57	69.51	6.50	64.79	76.10	14.10
2022	AGPCNet[42]	84.25	88.76	5.20	59.07	71.77	5.86	74.02	79.02	9.36
2022	DNANet[25]	90.14	93.34	5.18	**65.82**	**74.41**	**4.90**	**78.46**	**80.13**	**9.45**
2023	LW-IRSTNet[28]	73.02	79.33	7.90	56.51	71.79	19.90	63.47	74.00	13.26
2023	**MDAFNet-48**	**93.42**	**95.58**	**1.89**	64.25	76.66	6.11	78.35	84.06	11.22

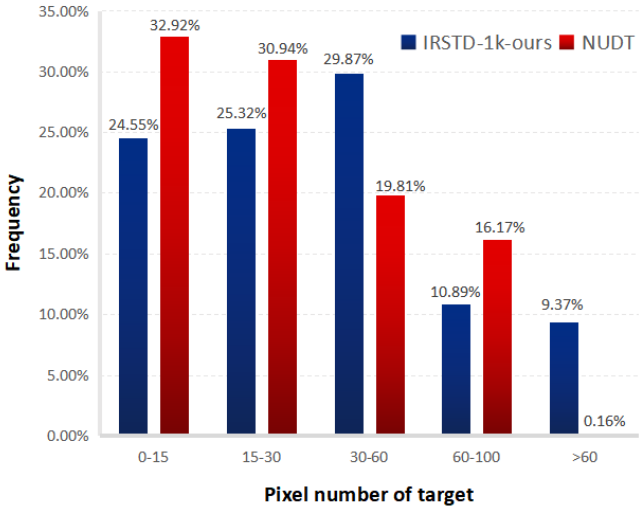

Fig. 4. Statistics on the target size of the NUDT and IRSTD-1k-ours datasets.

1) Normalized Intersection Over Union: nIoU is defined as

$$\text{nIoU} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{TP}[i]}{T[i]+P[i]-\text{TP}[i]}\right) \tag{9}$$

where N is the total number of samples, TP[·] denotes the number of true-positive pixels, and $T[\cdot]$ and $P[\cdot]$ denote the number of ground truth and predicted positive pixels, respectively.

2) Probability of Detection (P_d): P_d measures ratio of correctly predicted target pixels N_{pred} over all the target pixels N_{targets}

$$P_d = N_{\text{pred}}/N_{\text{targets}} \tag{10}$$

3) False-Alarm Rate (F_a): F_a measures the ratio of false predicted target pixels N_{false} over all the pixels N_{all} in the image

$$F_a = N_{\text{false}}/N_{\text{all}} \tag{11}$$

TABLE III

COMPARISON OF METRICS RELATED TO RUNNING SPEED OF DIFFERENT ALGORITHMS

Method	FOLPs(G)	Params(M)	FPS
UNet	54.73	31.03	125
ALCNet	3.70	0.37	83
LSPM	61.70	31.14	109
ACM-FPN	0.28	0.39	168
LW-IRSTNet	0.30	0.16	59
AGPCNet	43.18	12.36	12
DNANet	14.28	4.70	27
MDAFNet-48	100.46	10.07	45

In this article, we used datasets of NUDT, IRSTD-1k, and IRSTD-1k-ours. The training and validation sets were randomly split with 1:1 ratio. The training was conducted for 200 epochs using the Adam [43] optimizer with an initial learning rate of 0.001, and the learning rate decay strategy adopts a stair-step decay, with the learning rate halved every 20 epochs. In both the comparison experiments and ablation study, the start channel number of our network is set to 48. Moreover, all the models are conducted in PyTorch [44] on a computer with an Intel[1] Core[2] i9-10980XE CPU at 3.00-GHz and an NVIDIA GeForce RTX3080 GPU.

B. Comparison Experiments

1) Quantitative Results: To demonstrate the superiority of our method, we compare our MDAFNet with several state-of-the-art (SOTA) traditional methods and CNN-based methods. The traditional methods include NEW_TOPHAT [27], IPI [15], PSTNN [16], FKRW [38], and MDWCM [39], while the CNN-based methods include UNet [24], ALCNet [40], LSPM [41], ACMNet [30], AGPCNet [42], DNANet [25],

[1]Registered trademark.
[2]Trademarked.

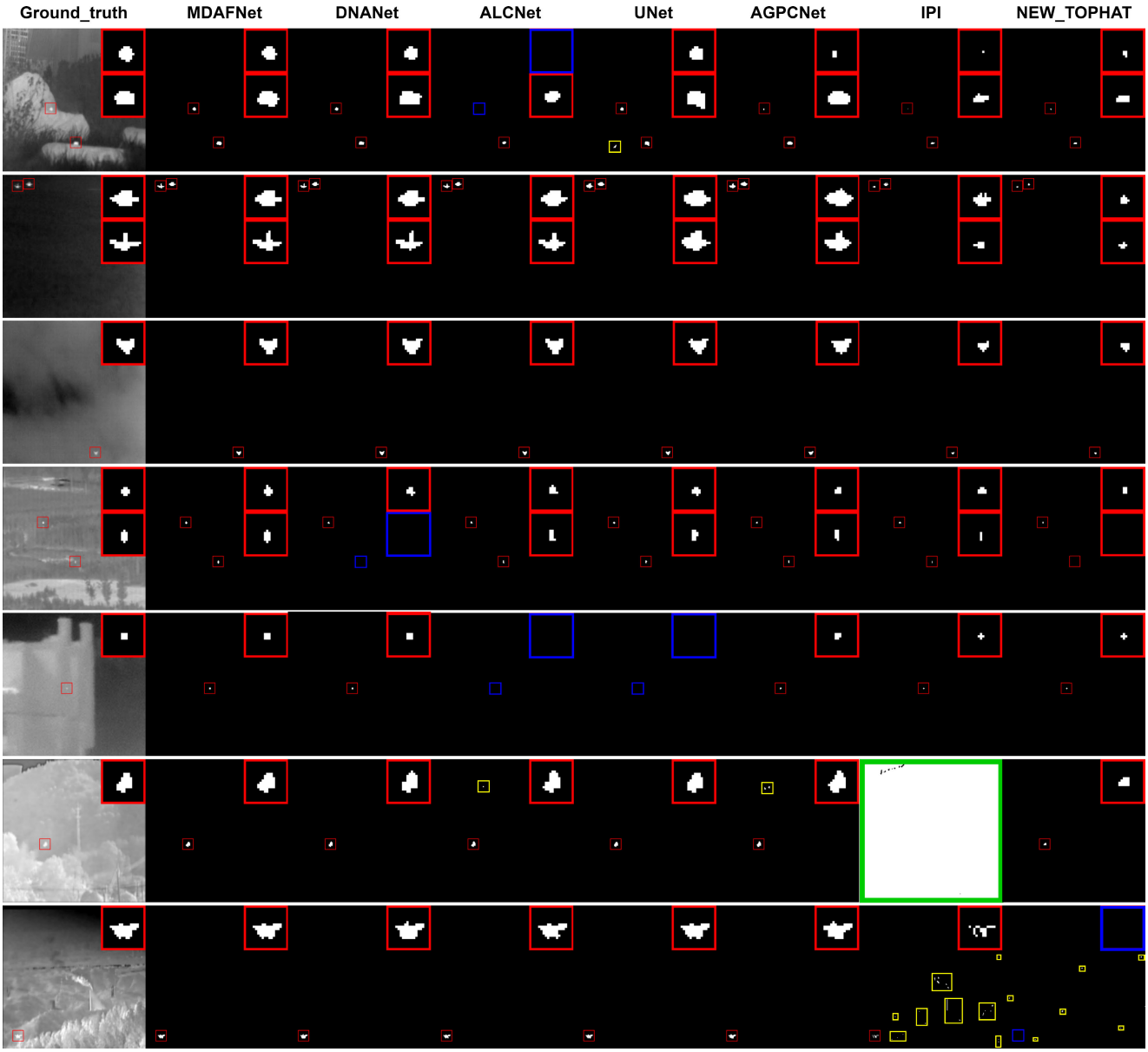

Fig. 5. Visualization of network output results: red boxes indicate the detected target regions, blue boxes represent missed detections, yellow boxes denote false alarms, and green boxes signify that the algorithm is not suitable for detecting objects in this image.

TABLE IV
ABLATION EXPERIMENT OF DETECTION HEAD ON BACKBONE OF BOTH BASIC BLOCK AND MDAF MODULE

Backbone		Detection Head		IRSTD-1k			NUDT		
MDAF	Basic	CF	FCN	nIoU(10^2)	Pd(10^2)	Fa(10^5)	nIoU(10^2)	Pd(10^2)	Fa(10^5)
×	✓	×	✓	72.58	79.71	11.76	87.74	91.08	3.62
✓	×	×	✓	73.56	78.42	18.82	88.80	91.74	4.87
×	✓	✓	×	76.95	81.72	11.14	90.79	94.02	4.21
✓	×	✓	×	78.35	84.06	11.22	93.42	95.58	1.89

LW-IRSTNet [28], and ours. To ensure fairness and persuasiveness in comparison experiments, the optimizer, learning rate, loss function, data augmentation, and other relevant parameters were configured based on the settings in each method's original article.

The quantitative evaluation results are presented in Table II. The improvements achieved by our MDAFNet over traditional methods are significant. This is due to the fact that traditional algorithms rely heavily on handcraft physical features, which limits them only to recognize images in certain scenarios, resulting in poor generalization performance. In contrast, CNN-based methods can learn discriminative features that enable them to adopt to various types of targets in different scenarios.

Among CNN-based methods, our MDAFNet-48 achieves the best performance on the NUDT dataset. Our network

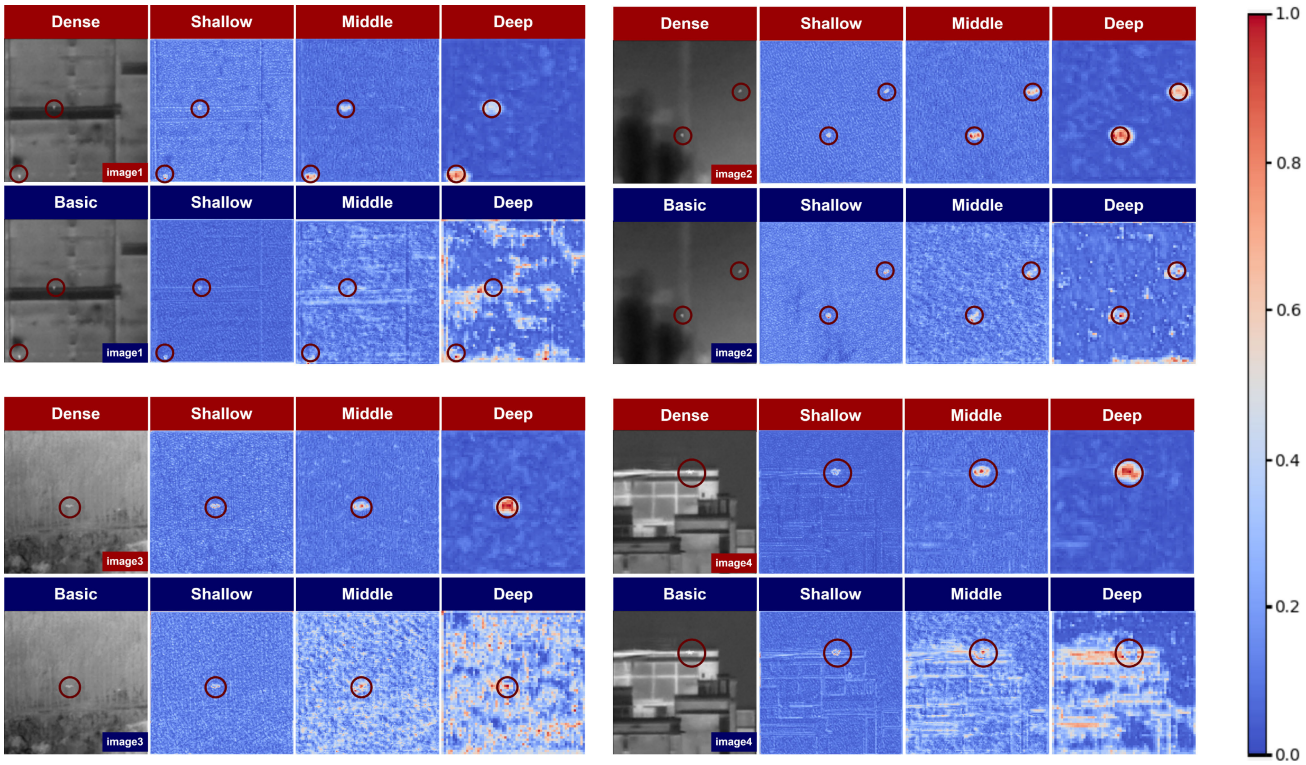

Fig. 6. Illustration of the feature maps in our network when using the backbone of both MDAF module and basic block. Shallow, middle, and deep correspond to S_1, S_2, and S_3 feature maps. These feature maps are obtained by summing up all the channels and normalizing them.

shows a significant improvement in normalized intersection over union (nIoU), P_d, and F_a values on the NUDT dataset compared with DNANet, with an increase of 3.28% and 2.24% and a decrease of 3.29×10^{-5}, respectively. However, on the IRSTD-1k and IRSTD-1k-ours datasets, the performance of nIoU is 1.57% and 0.06% lower than that of DNANet. This is because of the deeper network structure of DNANet, which includes five downsampling operations and is more suitable for detecting larger sized targets. Our network is shallower with fewer downsampling operations, making it more suitable for detecting smaller sized targets. As shown in Fig. 4, the NUDT dataset has a significantly higher number of small-sized targets compared with IRSTD-1k-ours. Therefore, our network exhibits superior performance on the NUDT dataset.

Furthermore, we compare the parameter count (Params), computational complexity (FLOPs), and interference speed (FPS) among different CNN-based algorithms, as shown in Table III. Although our model has a larger number of parameters and computations than DNANet and AGPCNet, our FPS is higher than them. The reason for the large number of parameters and computations is that our network is wider. And there are two reasons for the faster inference speed. First, our neural network consists of only 39 convolution operations, which is significantly fewer than those in DNAnet and AGPCNet. Due to the parallel computing capabilities of GPUs for each convolutional operation, our network achieves faster computational speed. Second, DNANet and AGPCNet incorporate many channel attention and spatial attention module compared with our network, which makes their interference time longer.

We also observe that all the comparative algorithms perform better on the modified IRSTD-1k-ours dataset compared with the IRSTD-1k dataset. This is because the modified dataset has higher annotation quality, resulting in more robust detection results.

2) Qualitative Results: The visualization results of the qualitative indicators are shown in Fig. 5. It can be observed that traditional algorithms are able to detect point targets with high SCR in simple backgrounds but tend to wrong detection in complex backgrounds. Moreover, traditional algorithms have poor adaptability to scene changes. When the features used by traditional algorithms are not prominent in backgrounds, the detection performance can be extremely poor, as illustrated in the sixth row and seventh column of Fig. 5. The diverse backgrounds in this image result in weak nonlocal correlations, rendering the IPI algorithm ineffective. Among CNN-based methods, it can be observed that our method achieves the best detection performance. Regardless of grassland, sky, complex terrain, urban buildings, or mountainous backgrounds, MDAFNet exhibits fewer false alarms and missed detections compared with other networks. This is because our network can effectively extract high-level semantic features of small targets in deep layers and CFHead demonstrates excellent adaptability to segment contours of targets of various sizes.

C. Ablation Study

1) Impact of Moderately Dense Adaptive Feature Fusion Module: To validate the effectiveness of our MDAF module in extracting features of small infrared targets, we replace the backbone with MDAF module and basic block, and the results are shown in Table IV. With the detection head of the network fixed, all the metrics of the MDAF module backbone are better than those of the basic block. There are

TABLE V

COMPARATIVE EXPERIMENT BETWEEN FCNHEAD AND CFHEAD ON UNET, ALCNET, DNANET, AND OURS. THE RED ARROW REPRESENTS THE PERCENTAGE INCREASE IN THE CORRESPONDING METRIC FOR CFHEAD COMPARED WITH FCNHEAD

Methods	Head	IRSTD-1k-ours			NUDT		
		nIoU(10^2)	Pd(10^2)	Fa(10^5)	nIoU(10^2)	Pd(10^2)	Fa(10^5)
UNet	FCNHead	69.05	75.24	16.19	88.53	91.70	1.87
	CFHead	72.05↑**3.00**	77.84↑**2.60**	10.28	88.78↑**0.25**	91.92↑**0.22**	2.23
ALCNet	FCNHead	73.40	76.65	7.78	82.57	87.29	3.07
	CFHead	75.44↑**2.04**	82.11↑**5.46**	11.37	89.26↑**6.69**	91.97↑**4.68**	2.17
DNANet	FCNHead	78.46	80.13	9.45	90.14	93.34	5.18
	CFHead	80.07↑**1.66**	82.21↑**2.08**	8.09	92.46↑**2.32**	95.07↑**1.73**	3.5
Ours	FCNHead	73.56	78.42	18.82	88.80	91.74	4.87
	CFHead	78.35↑**4.81**	84.06↑**3.95**	11.22	93.42↑**2.30**	95.58↑**2.12**	1.89

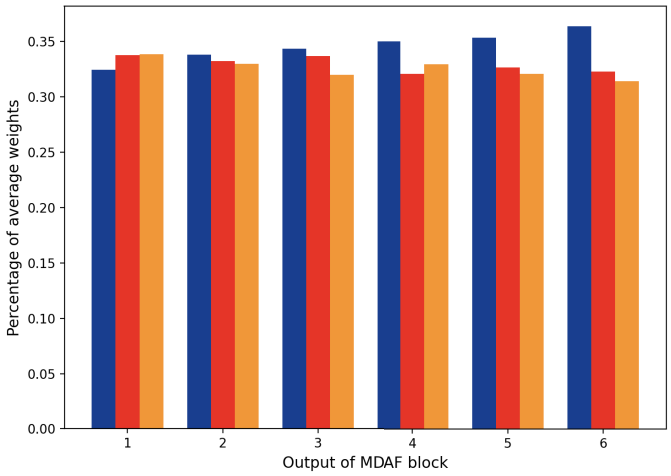

Fig. 7. Horizontal axis represents the first, second, and third layer from six MDAF modules, and the vertical axis represents the proportion of the absolute sum of the 1×1 convolution kernels applied to the ith$(i = 1, 2, 3)$ layer to the absolute sum of the 1×1 convolution kernels applied to all the layers within every MDAF module.

three reasons: First, dense connection within MDAF module enables the production of the last layer be guided by all the previous layers, strengthening feature propagation, and thus making network maintain the small target information in deep layers for better performance. As illustrated in Fig. 6, we compare the visualization results of S_1, S_2, and S_3 feature maps when using MDAF module and basic block as the backbone. We find that features of infrared small targets are overwhelmed in the deeper layers with basic block, while the MDAF module effectively extracts their semantic information. Second, the output of the basic block is simply the sum of the first and last layer features ignoring the second feature maps, but, as shown in Fig. 7, the sum of the 1×1 convolutional coefficients on the second-level feature maps almost accounts for one-third of the total output, indicating that second-level feature maps are also important and our MDAF module enhances the reuse of them. Third, the output of the basic block does not focus on which layer's information is most useful while the MDAF module can adaptively allocate appropriate weights to each layer through learning 1×1 convolution kernels, as illustrated in Fig. 7, As the network depth increases, the relative importance of the features in the first layer of the MDAF module gradually increases. This proves that the MDAF module increases the flexibility of network feature fusion and enables the network to extract more efficient feature information.

2) Impact of CFHead: To demonstrate the benefits of introducing our proposed CFHead, we compare the detection results of FCNHead and CFHead on UNet, ALCNet, DNANet, and our network. According to the results presented in Table V, both on the NUDT and IRSTD-1k-ours datasets, UNet, ALCNet, DNANet, and our methods all exhibit significant performance improvements. If the CFHead is replaced by FCNHead, the performance suffers decreases of 4.18% and 3.95% and an increase of 7.60 $\times$ 10^{-5} in terms of nIoU, P_d, and F_a values for MDAFNet on the IRSTD-1k-ours dataset. This is because our proposed CFHead introduces an auxiliary loss before generating the output, effectively minimizes the number of erroneous pixels in the coarse result, and then convolution with different dilation rates further refines the contour of the target resulting in a more precise estimation of the object boundaries and shapes. These outcomes also signify that the plug-and-play CFHead is not only applicable to our proposed MDAFNet but also demonstrates strong generalization performance when applied to other infrared target segmentation networks.

3) Impact of Multicontrast Under Different Brightness Data Augmentation Method: During the training process, data augmentation is used to increase the number of the dataset, enhance the model's generalization capability, and mitigate overfitting. We first compare the performance of MDAFNet before and after data augmentation on the IRSTD-1k-ours and NUDT datasets and find that with our multicontrast data augmentation method, the validation set has significantly improved performance. This indicates that the proposed data augmentation method effectively reduces network overfitting and improves the generalization of the network. To further demonstrate the importance of our proposed data augmentation method, we also compare the performance of ALCNet and LW-IRSTDNet before and after using our data augmentation method and find that all the metrics' values show a significant improvement. This further demonstrates the applicability of our proposed data augmentation method for detecting

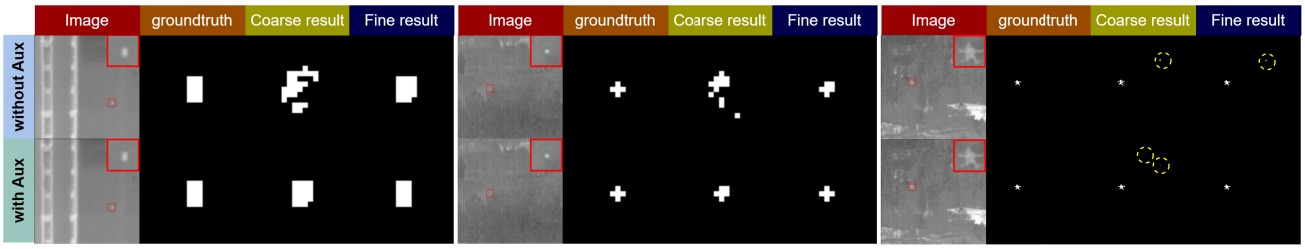

Fig. 8. Illustration of the comparison results of coarse and fine results before and after adding auxiliary losses.

TABLE VI

ABLATION EXPERIMENT OF MULTICONTRAST UNDER DIFFERENT BRIGHTNESS DATA AUGMENTATION METHOD ON ALNET, LW-IRSTDNET, AND OURS. THE RED ARROW REPRESENTS THE PERCENTAGE INCREASE IN THE CORRESPONDING METRIC WHEN USING OUR DATA AUGMENTATION METHOD

Method	Data_aug	IRSTD-1k-ours			NUDT		
		nIoU(10^2)	Pd(10^2)	Fa(10^5)	nIoU(10^2)	Pd(10^2)	Fa(10^5)
ALCNet	×	73.40	76.65	7.78	82.57	87.29	3.07
	✓	75.13↑**1.73**	79.22↑**2.57**	9.99	90.14↑**7.57**	93.15↑**5.86**	1.47
LW-IRSTNet	×	63.67	74.00	13.26	73.02	79.33	7.90
	✓	64.31↑**0.64**	76.67↑**2.67**	17.46	76.18↑**3.16**	81.15↑**1.82**	7.00
Ours	×	76.90	80.11	8.57	91.12	93.46	2.05
	✓	78.35↑**1.47**	84.06↑**3.95**	11.22	93.42↑**2.30**	95.58↑**2.12**	1.89

TABLE VII

IMPACT OF LOSS WEIGHT α ON DETECTION PERFORMANCE

α	IRSTD-1k-ours			NUDT		
	nIoU(10^2)	Pd(10^2)	Fa(10^5)	nIoU(10^2)	Pd(10^2)	Fa(10^5)
0	76.76	78.72	1.85	93.03	95.41	9.04
0.2	75.95	82.19	12.61	93.08	95.63	2.00
0.4	77.98	81.68	10.92	93.30	95.40	1.93
0.6	78.20	81.31	8.60	92.89	95.66	2.16
0.8	**78.35**	**84.06**	**11.22**	**93.42**	**95.58**	**1.89**
1	77.04	82.26	13.51	93.10	95.84	2.67

infrared small targets. The quantitative results are shown in Table VI.

D. Parameters Sensitivity Study

1) Impact of Auxiliary Loss: To investigate the impact of auxiliary loss on detection performance, we compare the detection results under various auxiliary loss weights α on MDAFNet, and the results are presented in Table VII. We find that the best detection performance is achieved at $\alpha = 0.8$. After adding the auxiliary loss, both the coarse and fine results are improved. This is because auxiliary losses can allow the network to generate more precise coarse predictions before producing the final prediction results and then modify it through convolution with different delition rates based on the coarse prediction. The quantitative metric results are presented in Table VIII, and the visualized results are shown in Fig. 8. By visualizing the results on the NUDT dataset, we also observe that the number of erroneous pixels in fine result is less than that in coarse result, which further indicates that the convolutions with different delition rates added behind the network achieve fine-tuned correction of the coarse results and increase the accuracy of the network's predictions.

E. Number of Convolutional Layers in Our Network

We conduct a series of experiments to explore the influence of network width on detection performance. We set the start channels to 12, 24, 36, 48, and 60. The results are presented in Table IX, and we find that with the increase in network width k, nIoU, and P_d values show an upward trend. This is because a larger number of parameters allows the network to learn richer features. However, when k reaches 48, the impact of network width on detection performance reaches saturation, and even shows a decreasing trend on the IRSTD-1k-ours dataset. To balance the parameter quantity and detection performance, in this article, we selected MDAFNet-48 as the baseline for further exploration.

F. Limitation of Our Method

To comprehensively analyze the performance of our network, we conducted tests on a large dataset of infrared target images with diverse backgrounds. We observe that the detection results are unsatisfactory when the target and background had a high degree of similarity. Similar findings are observed when testing other CNN-based algorithms. The visualized detection results are shown in Fig. 9. This issue may be attributed to the presence of numerous noise points in the background that resemble the target, which makes it challenging for the neural network to extract semantic differences between the target and the background. To address this issue, it is possible to enhance the improvement of the accuracy of CNN-based methods by incorporating suitable attention mechanisms and implementing appropriate data augmentation methods in future research.

TABLE VIII

ABLATION EXPERIMENT OF AUXILIARY LOSS ON BOTH NUDT AND IRSTD-1K-OURS. THE RED ARROW REPRESENTS THE PERCENTAGE INCREASE AFTER ADDING THE AUXILIARY LOSS

Dataset	Aux_loss	Coarse_result			Fine_result		
		nIoU(10^2)	Pd(10^2)	Fa(10^5)	nIoU(10^2)	Pd(10^2)	Fa(10^5)
NUDT	×	22.18	51.55	63.57	93.03	95.41	1.85
	✓	83.83↑**61.65**	95.21↑**43.66**	8.81	93.42↑**0.39**	95.58↑**0.17**	1.88
IRSTD-1k-ours	×	72.48	78.13	14.82	76.76	78.72	9.04
	✓	78.35↑**5.77**	84.38↑**6.25**	11.80	78.35↑**1.59**	84.06↑**5.34**	11.22

TABLE IX

IMPACT OF NETWORK WIDTH k ON DETECTION PERFORMANCE

k	IRSTD-1k-ours			NUDT		
	nIoU(10^2)	Pd(10^2)	Fa(10^5)	nIoU(10^2)	Pd(10^2)	Fa(10^5)
12	69.77	76.07	10.50	88.52	90.87	2.54
24	73.49	82.00	14.90	91.11	93.89	2.62
36	75.18	76.50	8.86	92.83	94.90	1.90
48	**78.35**	**84.06**	**11.22**	**93.42**	**95.58**	**1.89**
60	77.73	83.74	15.39	93.10	95.84	2.67

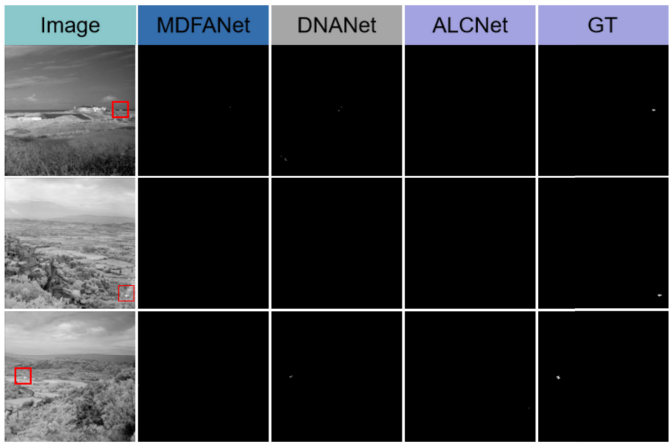

Fig. 9. Visualization of limitations in CNN-based methods in specific scenarios. The red box on the image indicates the target region, and the "GT" represents for the ground-truth label.

V. CONCLUSION

In this article, we propose an MDAFNet to achieve SIRST detection. Different from other CNN-based detection method, we explicitly handle the problem of infrared small targets being lost in deep layers with an MDAF module. We also design a CFHead and introduce auxiliary loss to achieve accurate prediction of the contours of small targets. Moreover, we develop a multicontrast under different brightness data augmentation method, which effectively improved generalization performance of the network. Experiments on the NUDT, IRSTD-1k, and IRSTD-1k-ours datasets have shown the superiority of our method over other SOTA methods. However, our analysis has revealed that the performance of network is adversely affected in scenarios where there is significant similarity between the target and the background. This issue is not specific to our approach but is a common challenge shared by other CNN-based methods. In future research, it is expected that the detection accuracy of the network can be improved using appropriate attention mechanisms and data augmentation methods.

REFERENCES

[1] A. Karim and J. Y. Andersson, "Infrared detectors: Advances, challenges and new technologies," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 51, Dec. 2013, Art. no. 012001.
[2] L. Becker, "Influence of ir sensor technology on the military and civil defense," in *Quantum Sensing and Nanophotonic Devices III*. Bellingham, WA, USA: SPIE, 2006, p. 61270S.
[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] M. Teutsch and W. Krüger, "Classification of small boats in infrared images for maritime surveillance," in *Proc. Int. WaterSide Secur. Conf.*, Nov. 2010, pp. 1–7.
[5] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 867–876.
[6] 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.
[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] 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.
[9] 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.
[10] 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.
[11] Y. Chen, B. Song, D. Wang, and L. Guo, "An effective infrared small target detection method based on the human visual attention," *Infr. Phys. Technol.*, vol. 95, pp. 128–135, Dec. 2018.
[12] 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.
[13] R. Kou, C. Wang, Q. Fu, Y. Yu, and D. Zhang, "Infrared small target detection based on the improved density peak global search and human visual local contrast mechanism," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6144–6157, 2022.
[14] J. Rivest, "Detection of dim targets in digital infrared imagery by morphological image processing," *Opt. Eng.*, vol. 35, no. 7, p. 1886, Jul. 1996.
[15] 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.
[16] 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.
[17] 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.
[18] H. Zhu, H. Ni, S. Liu, G. Xu, and L. Deng, "TNLRS: Target-aware non-local low-rank modeling with saliency filtering regularization for infrared small target detection," *IEEE Trans. Image Process.*, vol. 29, pp. 9546–9558, 2020.
[19] Z. Zhou, J. Feng, X. Wu, J. Shi, and X. Zhang, "Spectral constrained residual attention network for hyperspectral pansharpening," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 2386–2389.

[20] J. Feng, N. Zhao, R. Shang, X. Zhang, and L. Jiao, "Self-supervised divide-and-conquer generative adversarial network for classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536517.

[21] J. Feng, Y. Liang, X. Zhang, J. Zhang, and L. Jiao, "SDANet: Semantic-embedded density adaptive network for moving vehicle detection in satellite videos," *IEEE Trans. Image Process.*, vol. 32, pp. 1788–1801, 2023.

[22] Y. Liang, J. Feng, X. Zhang, J. Zhang, and L. Jiao, "MidNet: An anchor-and-angle-free detector for oriented ship detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612113.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[24] 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.* (Lecture Notes in Computer Science), vol. 9351, 2015, pp. 234–241.

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

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] X. Bai and F. Zhou, "Analysis of new top-hat transformation and the application for infrared dim small target detection," *Pattern Recognit.*, vol. 43, no. 6, pp. 2145–2156, Jun. 2010.

[28] R. Kou, C. Wang, F. Huang, Y. Yu, Z. Peng, and Q. Fu, "LW-IRSTNet: Lightweight infrared small target segmentation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5621313.

[29] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8508–8517.

[30] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," 2020, *arXiv:2009.14530*.

[31] A. Galdran et al., "Data-driven color augmentation techniques for deep skin image analysis," 2017, *arXiv:1703.03702*.

[32] J. Jin, A. Dundar, and E. Culurciello, "Robust convolutional neural networks under adversarial noise," 2015, *arXiv:1511.06306*.

[33] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.

[34] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[35] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[36] H. Wu, S. Zheng, J. Zhang, and K. Huang, "GP-GAN: Towards realistic high-resolution image blending," 2017, *arXiv:1703.07195*.

[37] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2016, pp. 234–244.

[38] Y. Qin, L. Bruzzone, C. Gao, and B. Li, "Infrared small target detection based on facet kernel and random walker," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7104–7118, Sep. 2019.

[39] R. Lu, X. Yang, W. Li, J. Fan, D. Li, and X. Jing, "Robust infrared small target detection via multidirectional derivative-based weighted contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[40] 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.

[41] L. Huang, S. Dai, T. Huang, X. Huang, and H. Wang, "Infrared small target segmentation with multiscale feature representation," *Infr. Phys. Technol.*, vol. 116, Aug. 2021, Art. no. 103755.

[42] 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.

[43] J. Duchi, E. Hazan, and Y. and, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Jul. 2011.

[44] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, vol. 32, May 2019, pp. 8026–8037.

Chengyu Li was born in Nanyang, China, in 2001. He received the B.E. degree in electronic information engineering from Tiangong University (TGU), Tianjin, China, in 2022. He is currently pursuing the M.E. degree with the National Key Laboratory of Science and Technology on ATR, College of Electronic Science and Technology, National University of Defense Technology (NUDT), Changsha, China.

His research interests include deep learning, automatic target recognition, infrared image processing, and remote sensing target recognition.

Yan Zhang was born in Shandong, China, in 1975. She received the B.E., M.E., and Ph.D. degrees from the National University of Defense Technology (NUDT), Changsha, China, in 1997, 2001, and 2008, respectively.

She is currently a Professor with the National Key Laboratory of Science and Technology on ATR, NUDT. Her research interests include infrared image processing, infrared polarization imaging, automatic target recognition, and target tracking.

Zhiguang Shi was born in Shandong, China, in 1975. He received the B.E. degree in automatic control from Shijiazhuang Mechanical Engineering College, Shijiazhuang, China, in 1996, and the M.E. and Ph.D. degrees in information and telecommunication systems from the National Key Laboratory of Science and Technology on ATR, College of Electronic Science and Technology, National University of Defense Technology (NUDT), Changsha, China, in 2002 and 2007, respectively.

He is currently an Associate Professor with NUDT. His research interests include clutter modeling, compressed sensing, automatic target recognition, and statistical analysis.

Yu Zhang was born in Jinan, China, in 1995. He received the B.E. degree in electronic information engineering from Shandong University (SDU), Jinan, in 2017, and the M.E. and Ph.D. degrees from the National University of Defense Technology (NUDT), Changsha, China, in 2019 and 2023, respectively.

His research interests include deep learning, automatic target recognition, infrared image processing, and remote sensing target recognition.

Yi Zhang was born in Changsha, China, in 1995. He received the B.E. degree in electronic information engineering from the Naval University of Engineering (NUE), Wuhan, China, in 2017, and the M.E. degree from the National University of Defense Technology (NUDT), Changsha, China, in 2023.

His research interests include image processing and automatic target recognition.