Context-Guided Reverse Attention Network With Multiscale Aggregation for Infrared Small Target Detection

Shunshun Zhong, Fan Zhang, and Ji’an Duan

***Abstract*—Infrared small target detection is one of the vital tasks in various infrared detection applications and has some typical challenges, such as small and dim target, background noise, and complex scenes. To address the problem, a context-guided reverse attention network is proposed to detect infrared small target by introducing context-guided module (CGM), multiscale aggregation block (MAB), and reverse attention module (RAM). The CGM is designed to capture the inherent property of semantic information from multiscale encode layer in pixel-level recognition. In order to eliminate the impact of low-level feature on computational complexity and ensure the detection performance, we design the MAB to aggregate multiscale feature. The RAM is integrated in decoder layer to combine the features from MAB and CGM for fusing the localization information and multiscale structural information. Extensive experiments on infrared small target datasets demonstrate that our method can achieve high detection accuracy and low false alarm rate compared with some state-of-the-art model-driven and data-driven methods.**

***Index Terms*—Deep learning, infrared image, neural network, target detection.**

I. Introduction

INFRARED segmentation small target detection (ISTD) plays a vital role in civil, industry, and military applications, such as autodriving, fire warning, leakage measurement, defect inspection, early warning system, missile tracking, and maritime surveillance [1]. Due to the long distance between target and detector, infrared small target often occupies a few pixels in a captured image, resulting in lacking shape and textures’ features [2]. In addition, owing to complex background turbulence, detector thermal noise disturbance, and optical scattering and diffraction interference, infrared image has low signal-to-noise ratio and low contrast between target and background, leading to targets easily to be overwhelmed by complex circumstance scenes [3]. Hereby, the ISTD is still a challenging task in the infrared search and tracking system.

Traditional methods to detect infrared small targets, for example, filter-based methods [4], local information-based methods [5], and data structure-based methods [6], often design handcrafted features to build filters or models fully based on prior knowledge. Therefore, this type of method has serious missed detections and is not feasible for all kinds of specific scenes.

Recently, deep learning methods divided into detection-based methods and segmentation-based methods have been developed for ISTD due to its powerful ability in feature extraction and feature representation [7], [8], [9]; this type of methods takes an infrared image as input and directly learns the small target features without introducing artificial prior and can be widely applied in the complex circumstance. However, the infrared dim and small targets are easy missed by multiple convolutions in deep learning framework, resulting in low target detection accuracy, especially for large field of view [10]. The attention mechanism can effectively employ information transferred from feature maps for identifying salient features [11] and is expected to preserve small target features in deep learning architecture.

In this study, an effective deep learning framework, context-guided reverse attention network (CgraNet), is proposed to detect infrared small target in complex background. The main contributions in this study are summarized as follows.

1) A CgraNet with multiscale aggregation is proposed to detect infrared small target in complex background.
2) The CGM, MAB, and RAM are introduced in our framework to capture local feature, surround context, and global context information, and aggregate localization information and multiscale structural information.
3) Comprehensive experiments on the public datasets show that our method can achieve high detection accuracy and low false alarm rate compared with some state-of-the-art methods.

II. Related Work

Recently, all kinds of methods have been proposed to detect infrared small targets in various backgrounds, and the mainstream methods can be divided into the traditional and deep learning methods according to whether artificial prior is adopted.

Manuscript received 7 July 2023; revised 18 August 2023 and 28 September 2023; accepted 9 October 2023. Date of publication 13 October 2023; date of current version 25 October 2023. This work was supported by the National Natural Science Foundation of China under Grant 52005520. *(Corresponding author: Fan Zhang.)*

The authors are with the School of Automation, Central South University, Changsha 410083, China, also with the State Key Laboratory of Precision Manufacturing for Extreme Service Performance, Central South University, Changsha 410083, China, and also with the Key Laboratory of Industrial Intelligence and Systems, Central South University, Changsha 410083, China (e-mail: 18373290071@163.com; zhangfan219@csu.edu.cn; zf20072225@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3324492

A. Traditional Methods

The filter-based methods, such as Top-Hat filter [12], Mexican-Hat filter [13], and bilateral filter [14], seek to estimate the background and enhance the target assuming that the targets are unrelated to the surrounding background and are treated as high-frequency components. However, in heterogeneous scenes, such as complex cluster and ground background, the high-frequency features of the target are no longer salient.

The local information-based methods, for example, local contrast measure (LCM) [15], human visual local contrast mechanism [16], and derivative dissimilarity measure [17], make full use of the local pixel difference of the target and surrounding background. However, for extremely faint small targets in highly complex real-world scenarios with considerable background interference, these local information-based methods may encounter serious misdetection, originating from the small contrast difference between the targets and background.

The data structure-based methods, for instance, infrared patch-image (IPI) model [18], total variation weighted low-rank constraint method [19], and kernel robust principal component analysis model [20], distinguish infrared targets from background according to their different structural features, such as the sparsity of the target and low rank of the background. Zhang et al. [21] designed a model based on the nonconvex rank approximation minimization (NRAM) joint norm to improve the detection ability of infrared small targets in complex background. Liu et al. [22] put a nonconvex tensor low-rank approximation method considering that different singular values have different importance and should be treated discriminatively. Nevertheless, the high computational overhead and the sensitivity of hyperparameters to changes in image scenes limit their applications in infrared tracking and searching.

B. Deep Learning Methods

The detection-based deep learning methods, such as regions with convolutional neural network (R-CNN) [23], faster R-CNN [24], single-shot multibox detector (SSD) [25], and you only look once (YOLO) [7], predict infrared small targets with anchor by employing the feature representation of convolutional neural network (CNN). Sommer et al. [26] designed faster R-CNN embedded region proposal network to form candidate regions for classifying and determining the infrared targets. Ding et al. [27] reconstructed SSD-based network by reducing low-resolution layers and enhancing high-resolution layer for further dropping the false alarm rate and increasing the precision. Mou et al. [28] improved YOLO model according to feature reassembly sampling for decreasing the loss of target features.

The segmentation-based deep learning methods for infrared small target detection can output pixel-level classification and target localization, which help deal with the model loss caused by the targets' small size. Wang et al. [29] used a conditional generative adversarial network (MDvsFA) within two generators and one discriminator to obtain high detection accuracy and low false alarm rate. Dai et al. [30] designed an asymmetric contextual modulation (ACM) framework to detect infrared small targets and embedded top-down and bottom-up modulation strategy for fusing high-level and low-level features. Huang et al. [31] established multiple local similarity pyramid modules (LSPMs) by leveraging attention mechanism to extract the multiscale features. Zhang et al. [32] present an attention-guided pyramid context network to compute local and global associations between the semantics of infrared small targets. Chuang et al. [33] designed a multiscale local contrast learning (MLCL) network to generate local contrast feature for infrared small target detection. Wang et al. [34] proposed a coarse-to-fine interior attention-aware network (IAAN) to detect infrared small target by considering that pixels from targets or backgrounds are correlated with each other. Li et al. [35] proposed a dense nested attention network to reduce the loss of targets by pooling layers in CNN. Wu et al. [36] constructed a single U-Net in U-Net framework with powerful generalization performance to learn the multilevel and multiscale feature representation. Wu et al. [37] developed a multilevel TransUNet (MTU) to adaptively extract long-distance features for space-based infrared detection. Pan et al. [38] proposed an attention with bilinear correlation network, including convolution linear fusion transformer, to enhance target feature and suppress noise. Kou et al. [39] developed a lightweight encoding and decoding structure to balance the computational efficiency and model accuracy for separating infrared small targets. Sun et al. [40] proposed a receptive-field and direction-induced attention network (RDIAN) to solve the interclass imbalance between targets and background by using the characteristics of target size and grayscale.

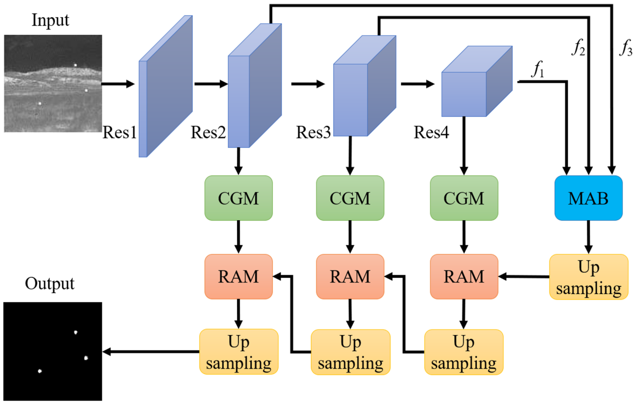

Fig. 1. Overview of the proposed framework CgraNet for infrared small target detection. CGM: context-guided module, RAM: reverse attention module, and MAB: multiscale aggregation block.

III. Methodology

A. Architecture Overview

Fig. 1 exhibits the architecture of designed CgraNet, which includes feature extraction stage, context-guided module (CGM), reverse attention module (RAM), multiscale aggregation block (MAB), and upsampling stage. The feature extraction stage, including Res1, Res2, Res3, and Res4 downsampling layers, is utilized to obtain the multiscale features of the infrared image and employs Res2Net [41] pretrained on ImageNet as the

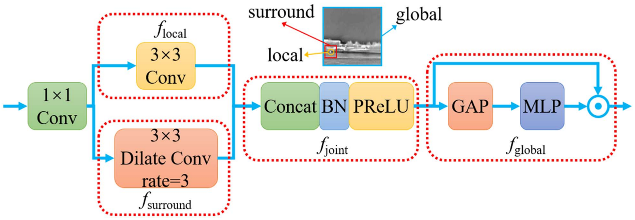

Fig. 2. Architecture of the CGM. PReLU: Parametric ReLU, GAP: global average pooling, MLP: multilayer perceptron. f_{local}: local feature extractor, f_{surround}: surround context extractor, f_{joint}: joint feature extractor, f_{global}: global context extractor, and ⊙: elementwise multiplication. The orange, red, and blue regions in infrared image correspond to local feature, surround context, and global context, respectively.

backbone of the CgraNet. The obtained multiscale feature maps from Res2, Res3, and Res4 layers are aggregated in high-level layers by using MAB to combine the contextual information and generate a global map as the initial guidance area for the subsequent upsampling stages. The CGM captures the essential property of semantic information in downsampling stages to learn the local feature and surrounding context. The RAM is designed to mine the boundary cues for building the relationship between target around areas and boundary cues. The low-level and high-level features are integrated into upsampling stages through RAM to accurately detect infrared small targets. Finally, a binary map of predicted target locations, whose size is the same as the input infrared image, is outputted from the end-to-end network CgraNet.

B. Context-Guided Module

The CGM is designed to capture the inherent property of semantic information from multiscale encode layer in pixel-level recognition based on human visual mechanism [42]. As shown in Fig. 2, it is difficult to categorize the infrared small target in complex background when the designed network only pays attention to the orange region within local information. If both the local information and the red region within the surround context information are considered, it is easier to distinguish target location from the background due to the fact that surround context with a larger area contains more useful information compared with the local region. Furthermore, the blue region contains the global context of the image scene, which can provide the global representation to recognize every pixel in the scene. Accordingly, the local feature, surround context, and global context may be considered in attention model to effectively improve the detection accuracy.

Hereby, the CGM is introduced in our framework to capture different region information and contains a local feature extractor (f_{local}), a surround context extractor (f_{surround}), a joint feature extractor (f_{joint}), and a global context extractor (f_{global}), as exhibited in Fig. 2. A 1×1 convolutional layer with slide of 1 is used at the entrance of CGM for reducing channel number to 1. f_{local} is constructed by a 3×3 standard convolutional layer with slide of 1 to learn the local feature from the eight neighboring feature vectors, which corresponds to the orange region. While, f_{surround} is constructed by a 3×3 dilated convolutional layer with dilation rate of 3 and a slide of 1 for obtaining a larger receptive field to learn the surround context, which corresponds to red region. f_{joint} is designed by a concatenation layer (Concat), the batch normalization (BN), and the Parametric ReLU (PReLU) operators to combine the local feature and surround context from the output of f_{local} and f_{surround}. In order to reinforce useful information and weaken useless ones in joint features, the global context can be treated as a weighted vector and channelwisely employed on the output of f_{joint}. Therefore, f_{global} is built by a global average pooling (GAP) layer and a multilayer perceptron (MLP) to aggregate and refine the global context in blue region of image. Finally, a scale layer is adopted to reweight the obtained global context on the joint feature.

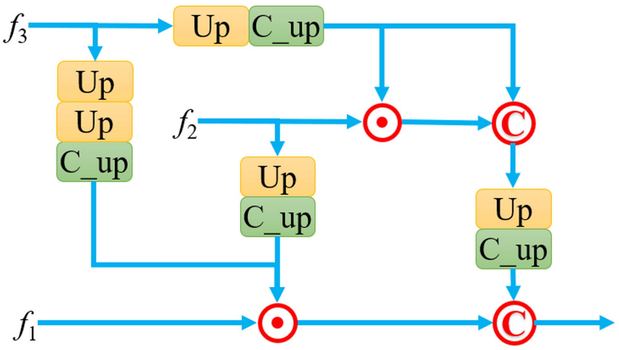

Fig. 3. Architecture of the MAB. Up: upsampling, C_up: conv_upsampling, ⊙: elementwise multiplication, and ©: concatenation.

C. Multiscale Aggregation Block

The encoder in the end-to-end framework can provide multilevel deep features, including high-level semantic feature and low-level spatial detail feature. The decoder combines these multilevel deep features to generate the accurate infrared small target feature maps. However, compared with high-level feature, the low-level feature contributes less to the detection performance of the framework. In addition, the low-level feature with large resolutions is integrated with high-level feature, resulting in increasing the computational complexity. Therefore, in order to eliminate the impact of low-level feature on computational complexity and ensure the detection performance, we designed a MAB in Fig. 3 to combine low-level and high-level contextual information. The original image with size of $H \times W \times C$ (H, W, and C represent the height, width, and channel, respectively) is inputted into Res2Net to generate different level features f_i ($i = 1, 2, 3, 4$), as shown in Fig. 1. The high-level features $\{f_1, f_2, f_3\}$ are fed into the MAB with a paralleled connection [43] to acquire a global map as the initial guidance area for the decoder.

D. Reverse Attention Module

The global map from the MAB can roughly locate the position of infrared target, while the CGM extracts multiscale feature within structural detail from the encoder layer. To aggregate the localization information and multiscale structural information,

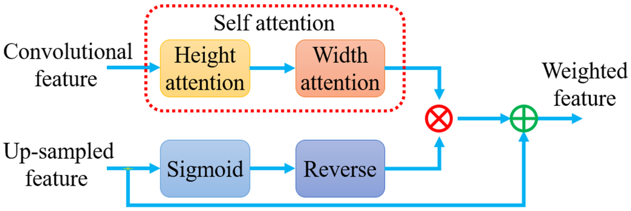

Fig. 4. Architecture of the RAM. ⊗: element vector–tensor multiplication and ⊕: element addition.

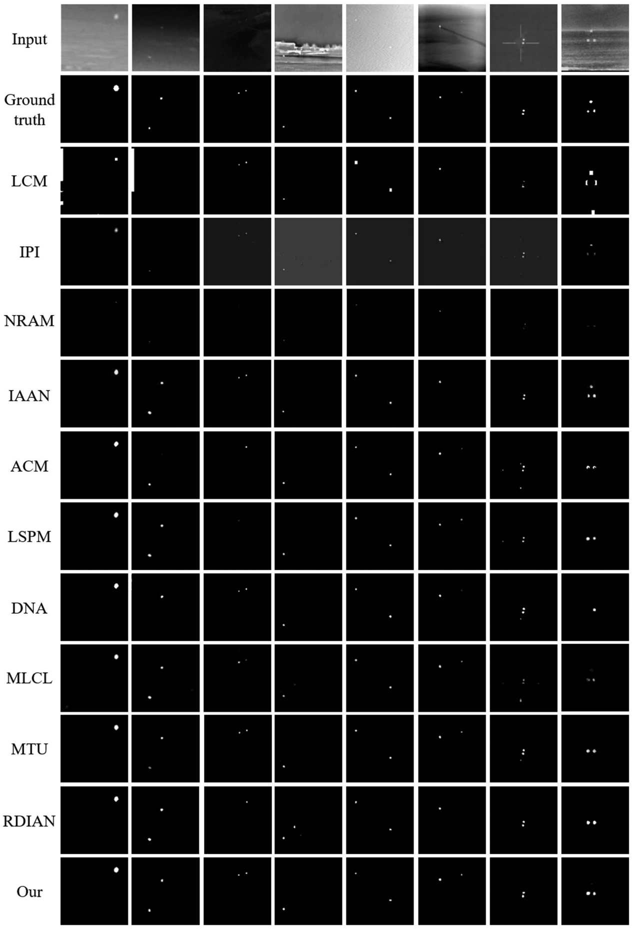

Fig. 5. Qualitative detection performance comparisons of our method with some state-of-the-art methods on NUST-SIRST dataset.

the RAM is designed to fuse convolutional feature and upsampled feature, as displayed in Fig. 4. The inputs of the RAM are multiscale feature target localization feature from the CGM and the MAB, respectively. The convolutional feature is inputted into a self-attention module to help learning global relationships. The attention operation maps a query and a set of key–value pairs as follows:

$$\text{attention}\,(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right) \tag{1}$$

where Q, K, V, and d_K represent the query, key, value, and dimension of key, respectively. Nevertheless, the self-attention may introduce huge computation complexity [44], especially for large spatial dimension input image. Hereby, the two-dimensional (2-D) attention is transformed into 1-D attention along the height and width axes to form height attention and width attention. In addition, the upsampled feature is treated by sigmoid activation function and reverse operation to erase the existing estimated regions, which refine the imprecise and coarse estimation into the accurate and complete prediction map. Therefore, the output of RAM is obtained by multiplying the output feature from self-attention module and reverse attention with residual upsampled feature [44], which is benefit for sequentially mining complementary regions and details.

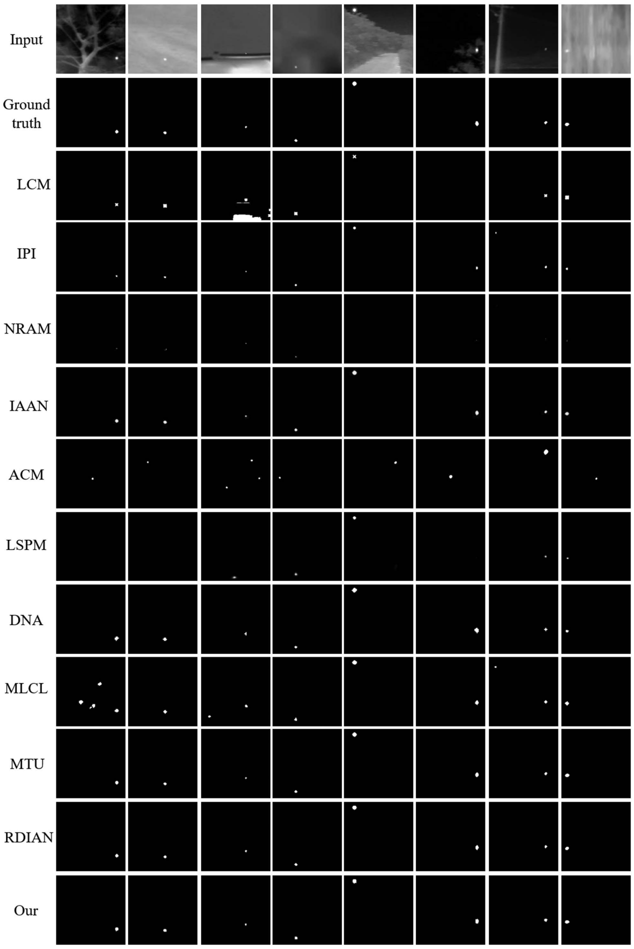

Fig. 6. Qualitative detection performance comparisons of our method with some state-of-the-art methods on NUDT-SIRST dataset.

E. Loss Function

In order to accurately calculate the global and local losses for model training, the intersection over union (IoU) (L_{IoU}) and binary cross entropy (BCE) (L_{BCE}) are weighted in unite loss function (L_{unite}), which can be written as [46] follows:

$$L_{\text{unite}} = \lambda_1 L_{\text{IoU}} + \lambda_2 L_{\text{BCE}} + \lambda_3 L_2 \tag{2}$$

$$L_{\text{IoU}} = 1 - \frac{|P \times \text{GT}|}{|P| + |\text{GT}| - |P \times \text{GT}|} \tag{3}$$

TABLE I
QUANTITATIVE PERFORMANCE COMPARISONS OF OUR METHOD WITH SOME STATE-OF-THE-ART METHODS ON NUST-SIRST DATASET

Methods	AUC	IoU	F	P_d(×10^{-2})	F_a(×10^{-6})	Time (s/200 images)	Parameter number (MB)
LCM	0.7648	0.1930	0.3096	5.56	717.71	2.52	/
IPI	0.8214	0.4064	0.5315	56.41	40.03	449.12	/
NRAM	0.8365	0.4195	0.5167	48.97	61.99	270.00	/
IAAN	0.9479	0.4601	0.6774	57.50	41.06	85.11	19.72
ACM	0.8634	0.4332	0.6733	48.68	33.84	7.45	0.39
LSPM	0.9427	0.4225	0.6419	52.84	46.66	130.02	31.58
DNA	0.8638	0.4495	0.6506	50.17	50.71	25.68	4.70
MLCL	0.9534	0.4296	0.6782	56.26	38.65	19.24	0.55
MTU	0.9420	0.4621	0.6737	55.32	45.91	7.34	12.75
RDIAN	0.8733	0.4076	0.6054	46.87	55.79	6.45	0.22
Our	0.9510	0.4766	0.6909	58.49	31.71	12.12	26.12

$$L_{\mathrm{BCE}} = -\frac{1}{N}\sum_{j\,=\,1}^{N}\left(\mathrm{GT}_j \log\left(P_j\right) + \left(1 - \mathrm{GT}_j\right)\log\left(1 - P_j\right)\right) \tag{4}$$

where P is the predicted result, and GT is the ground truth. N is the number of pixels in an image, and j represents the pixel index of the image. L_2 is the regularization term. λ_1, λ_2, and λ_3 are the learning hyperparameters. It is empirically set as $\lambda_1 = 2$, $\lambda_2 = 7$, and $\lambda_3 = 1$.

IV. EXPERIMENTS

A. Experimental Setting

The datasets used in the experiment are the public datasets NUST-SIRST [29] and NUDT-SIRST [35] to successively select 10 000 images for training and 200 images for testing. Considering that the dramatic scene changes in infrared small target detection, the testing images partly are selected from NUAA-SIRST dataset [30], which can well demonstrate model's generalization. For infrared small target detection in our research, applying conventional bounding box regression requires setting a low IoU threshold, which will increase the false alarm rate and position errors. Hence, we treat the detection task as a semantic segmentation to accurately estimate the performance of designed network and some other models. The experiment is conducted on a computer with 12G NVIDIA RTX 3060 GPU by using Python and PyTorch. The size of the input infrared image is 128×128, and the batch size is set to 12. The Adam optimizer with L_2 regularization and the "poly" learning rate policy with an initial learning rate of 10^{-4} and a power of 0.9 are employed. The receiver operating characteristic (ROC) and precision–recall (PR) curves are used to directly show detection performance. The evaluating metrics in the experiments are the area under the ROC curve (AUC), the IoU, and the F measure,

TABLE II
QUANTITATIVE PERFORMANCE COMPARISONS OF OUR METHOD WITH SOME STATE-OF-THE-ART METHODS ON NUDT-SIRST DATASET

Methods	AUC	IoU	F	P_d (×10^{-2})	F_a (×10^{-6})
LCM	0.7352	0.2691	0.3000	40.15	287.2
IPI	0.8017	0.3152	0.5926	90.74	104.0
NRAM	0.8462	0.394	0.5938	98.37	153.0
IAAN	0.9298	0.6186	0.7134	95.79	91.00
ACM	0.7164	0.2125	0.5442	4.12	96.20
LSPM	0.7141	0.2111	0.5651	48.02	32.11
DNA	0.9261	0.5860	0.7025	79.71	55.97
MLCL	0.9083	0.5836	0.6348	59.68	52.94
MTU	0.9297	0.6117	0.7129	94.99	113.1
RDIAN	0.9293	0.6198	0.7173	97.89	24.47
Our	0.9534	0.6245	0.7654	98.41	18.35

which can be deduced as follows [47]:

$$\text{IoU} = \frac{|P \times \text{GT}|}{|P| + |\text{GT}| - |P \times \text{GT}|} \tag{5}$$

$$F = \frac{(\beta^2 + 1) * P_d * F_a}{\beta^2 * (P_d + F_a)} \tag{6}$$

where P and GT are the predicted result and ground truth, respectively. P_d and F_a are successively the detection probability and false alarm rate, and parameter β is set to 1.

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

For qualitatively comparing segmentation performance, some representative infrared small target detection results of our designed network and some other models are exhibited in Figs. 5 and 6. The first and second rows successively correspond to the input images and ground truth, and the third to ninth rows represent the predicted results of LCM, IPI, NRAM, IAAN, ACM, LSPM, DNA, MLCL, MTU, RDIAN, and our method, respectively.

It can be seen in the second column of Fig. 5 that most methods fail to detect the dim small targets, specifically, the LCM, IPI, and ACM methods mistakenly detect two targets with incorrect locations due to the interference of cloudy clutter backgrounds. When facing general obstacles, such as high buildings with high brightness in the last column, our model can also locate small targets with accurate gray values. In addition, when detecting small targets with a low signal-to-clutter ratio in the first, third, and fifth columns, most methods have a high false alarm rate, and it is difficult to detect real targets with accurate quantity and brightness. Our model has achieved high detection accuracy and low false alarm rate in such challenging scenes, which is mainly attributed to the multiscale contextual feature fusion by the proposed MAB. It can be seen in the fourth column that our method can handle the input image with high boundary contrast of background and obtain a clearer target with exact contour. This is mainly due to the long-range global interactions of features built by the RAM and MAB, resulting in discriminating the spatial contours of small targets and edge information of the background [48]. When detecting multiple targets in the seventh column, our model can also process the situation with high detection accuracy compared with other methods. It can be seen in the eighth column that our method misses one target at the edge of the input image, which is the same to other methods. Therefore, the single-frame detection methods are often difficult to discriminate targets within edge region, and the multiframe detection methods may address the problem. Furthermore, compared with other methods, our method can produce output with precise target localization and shape segmentation under very low false alarm rate on NUDT-SIRST dataset in Fig. 6, which demonstrates the generalization of our method.

Figs. 7 and 8 illustrate the 3-D visual detection results on NUST-SIRST and NUDT-SIRST datasets for different methods and express that our method is more robust to these complex backgrounds. For example, in the fifth column of Fig. 7, the two targets are close to each other, which mistakes some models learning. However, our model can segment the targets well without introducing false alarm.

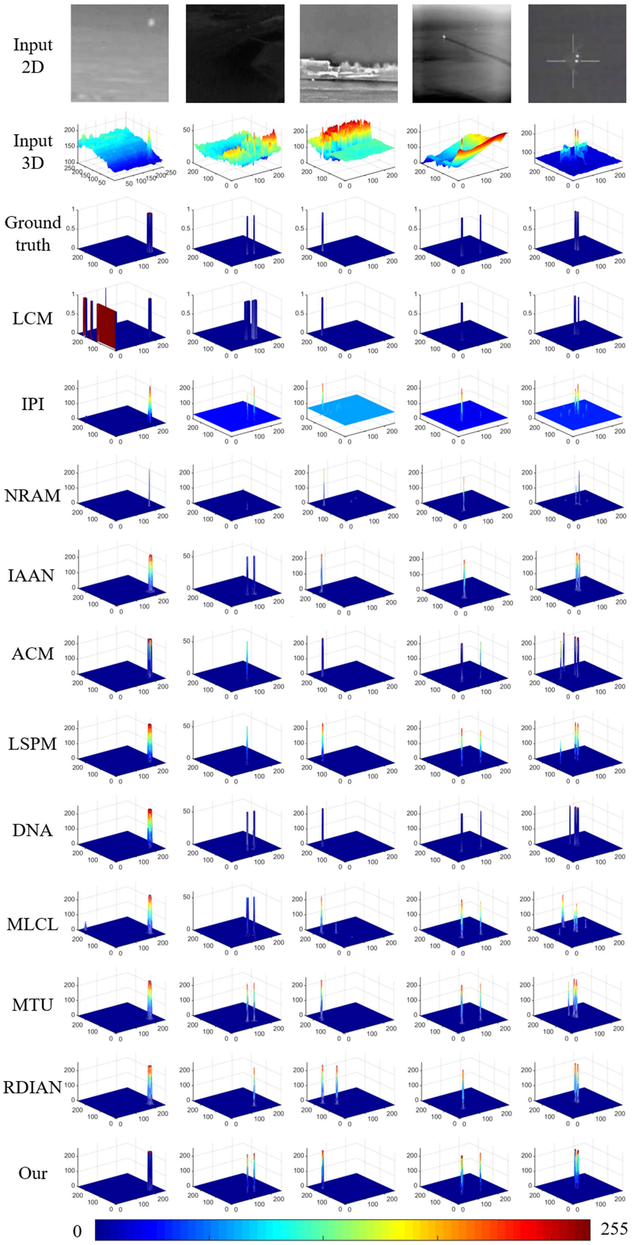

Fig. 7. Three-dimensional gray-level distributions of detection results of our method and some state-of-the-art methods on NUST-SIRST dataset.

Tables I and II list the quantitative estimation on NUST-SIRST and NUDT-SIRST datasets of our network and some other models, namely, contrast mechanism-based method LCM, low-rank sparse decomposition method IPI and NRAM, and typical data-driven methods, including IAAN, ACM, LSPM, DNA, MLCL, MTU, and RDIAN. It can be seen that our method has an obvious increase in AUC, IoU, and F compared with the conventional model-driven methods. This is because the conventional methods rely heavily on handcrafted features and have difficulty in adapting the variations of targets in complex backgrounds. Furthermore, compared with the data-driven deep

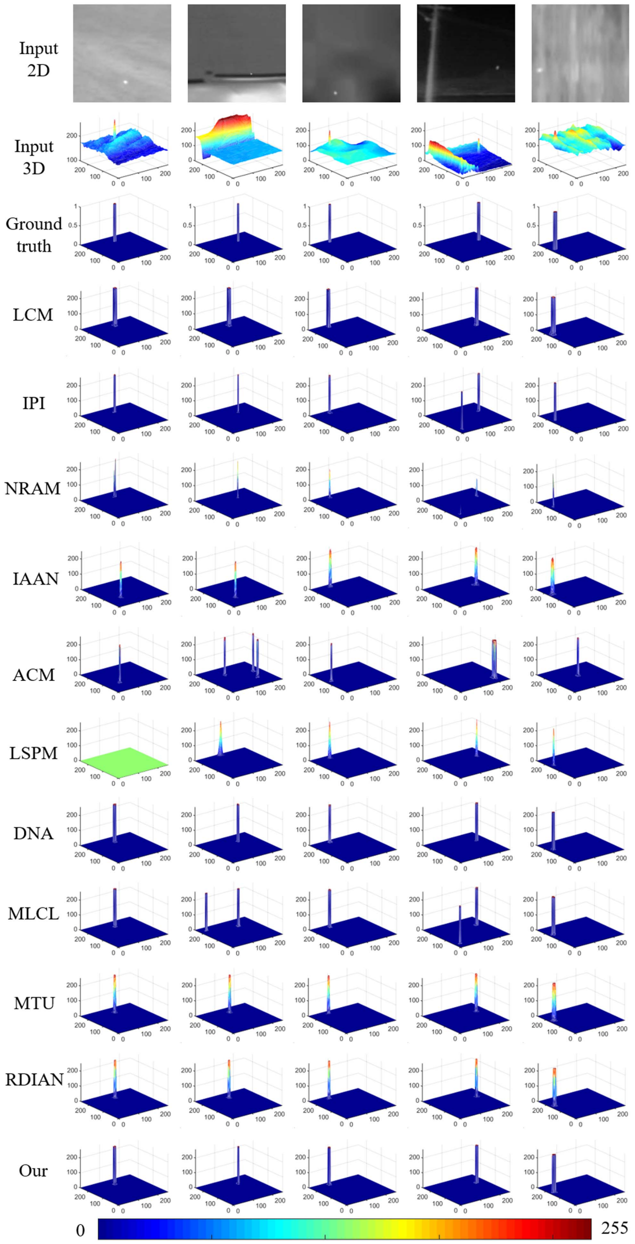

Fig. 8. Three-dimensional gray-level distributions of detection results of our method and some state-of-the-art methods on NUDT-SIRST dataset.

learning methods ACM, LSPM, MLCL, IAAN, DNA, MTU, and RDIAN, our method on NUST-SIRST dataset apparently increases the IoU value by 6.8%, 9.5%, 7.7%, 0.6%, 2.9%, 3.0%, and 14.4%, and increases the F value by 1.8%, 6.8%, 1.1%, 1.2%, 5.4%, 2.4%, and 12.3%, respectively. Our method also behaves competitive performance in AUC with value of 0.9510 for NUST-SIRST dataset compared with other deep learning methods. The metrics of AUC, IoU, and F for our method on NUDT-SIRST dataset reach 0.9534, 0.6245, and 0.7654, which are better than those of the compared methods. In addition, our model achieves high detection probability P_d and low false alarm rate F_a for both NUST-SIRST and NUDT-SIRST. These experimental results show that our network can preferably detect infrared small targets in complex backgrounds and segment targets more accurately than some other models. This is mainly attributed to the designed framework being able to adapt well to various challenges, such as cluttered background, target location, and small target, and thus achieving better performance.

TABLE III
EXPERIMENTAL RESULTS OF THE ABLATION STUDY FOR DIFFERENT LOSS FUNCTIONS ON NUST-SIRST DATASET

Loss function	AUC	IoU	F
IoU loss	0.9098	0.4580	0.5410
Dice loss	0.9433	0.3834	0.5407
Focal loss	0.9598	0.3813	0.6474
BCE	0.9713	0.3768	0.6001
IoU+Dice loss	0.9374	0.4127	0.5967
IoU+Focal loss	0.9435	0.4657	0.6178
Unite loss	0.9510	0.4766	0.6909

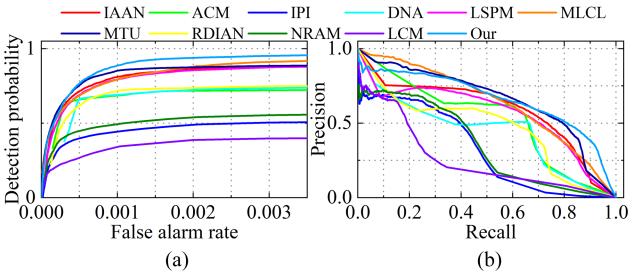

Fig. 9. (a) ROC and (b) PR curves of our method and some state-of-the-art methods on NUST-SIRST dataset.

Additionally, although the CGM, RAM, and MAB are added in our framework, the overall network parameter number is 26.12 M, which is less than that of the LSPM method. The interference time on the GPU of our method is only 12.12 s for testing 200 images, which is superior to that of most deep learning methods under the same device. Therefore, our method with high detection accuracy and high efficiency can be well applied in infrared guiding and tracking.

We further evaluate our framework and other models by using receiver operating characteristic (ROC) and PR metrics on NUST-SIRST dataset, as shown in Fig. 9. The ROC curve of our method is the closest to the top-left corner of Fig. 9(a). Compared with other methods, our method has a larger AUC, which expresses its high detection accuracy and low false alarm rate. Namely, our model can accurately segment infrared small targets and better suppress complex backgrounds simultaneously. The PR curve in Fig. 9(b) illustrates

TABLE IV
EXPERIMENTAL RESULTS OF THE ABLATION STUDY FOR DIFFERENT MODULES IN THE PROPOSED FRAMEWORK ON NUST-SIRST DATASET

CGM1	CGM2	CGM3	RAM1	RAM2	RAM3	MAB	AUC	IoU	F
×	×	×	×	×	×	×	0.8575	0.2436	0.4357
×	×	×	√	×	×	×	0.9057	0.3377	0.5024
√	×	×	√	×	×	×	0.9170	0.3649	0.5242
√	√	×	√	√	×	×	0.9211	0.4093	0.5647
√	√	√	√	√	√	×	0.9498	0.4130	0.6226
√	√	√	√	√	√	√	0.9510	0.4766	0.6909

that our method can focus on the target location in challenging scenarios for obtaining high precision and low recall, which is possibly due to the multiscale feature fusion following the attention mechanism for precisely capturing small targets in our framework.

C. Ablation Study

To verify the effectiveness of the designed unite loss function, IoU loss, Dice loss, focal loss, BCE loss, and their combinations are used as loss functions in the ablation experiment, and the experimental results on NUST-SIRST dataset are listed in Table III. Apparently, the IoU loss function can help the network achieve higher IoU values of 0.4580, while the network obtains lower AUC and *F* with values of 0.9098 and 0.5410. The IoU loss drives the maximal overlap between the ground truth and the predicted results, and jointly regresses all the segmentation variables as a whole unit, which avoids the issue of inaccurate measurement of pixel accuracy in the case of category imbalance between small target and background [49]. Therefore, more accurate predication and faster training convergence can be achieved by using the IoU loss. However, the IoU loss spreads more attention to the similarity of intersection areas between the predicted results and the ground truth, and easily confuses the small target in a complex background, which leads to a low *F* value. In addition, our network can obtain higher *F* values of 0.6474 when the focal loss function is used. Due to assigning greater weights to hard-to-classify examples compared with the easy-to-classify examples, the focal loss can well balance the small target and background during model training by setting a dynamic weight to adapt with learning accuracy. It can pay more attention on the small targets, which are easily misclassified due to class imbalance. When BCE loss is used to train our model, the highest AUC with value of 0.9713 is achieved since the BCE loss can minimize the oversmoothing impact of pixel loss while maintaining good deblurring effect [50]. Therefore, the unite loss function combining the IoU loss and BCE loss can acquire higher IoU and *F* values reaching 0.4766 and 0.6909, respectively.

To verify the validity of the CGM, RAM, and MAB in our proposed model, ablation experiments on NUST-SIRST dataset are conducted, and the corresponding results are listed in Table IV. CGM1, CGM2, and CGM3 guide the features from low level to high level in Fig. 1, and the same representation refers to RAM1, RAM2, and RAM3. Res2Net 50 is used as the baseline and obtains 0.8575 in AUC, 0.2436 in IoU, and 0.4357 in *F*. When the RAM1 module is added to the baseline, the AUC, IoU, and *F* values are obviously improved, which originates from the fusion of spatial details and semantic information to detect targets. When the CGM is added to the baseline with RAM, the segmentation performance further improves. Specifically, the AUC, IoU, and *F* values successively reach 0.9211, 0.4093, and 0.5647, as CGM1 and CGM2 are integrated into the Res2 and Res3 layers with RAM1 and RAM2. The better results are obtained by integrating all CGM and RAM into our framework and boost the segmentation performance by 10.7% in AUC, 69.5% in IoU, and 42.8% in *F* compared with the baseline. Finally, when all CGMs and RAMs with MAB are added to the framework, the best performance is achieved with the values of 0.951, 0.4766, and 0.6909 for AUC, IoU, and *F*, respectively. It is due to this that the target features of the downsampling operation can be enhanced by aggregating localization information and multiscale structural information [51].

The visual feature evolution is also illustrated in Fig. 10 to explore the effects of the CGM, RAM, and MAB. It can be observed that the feature maps from Res2 can focus on the target and edge background, which will cause interference to accurately learn small targets [51]. The CGM relaxes the edge background and maintains the small target, which improves the target detection performance. The feature maps from MAB behave the accurate target location, originating from the combination of low-level and high-level contextual information. The RAM aggregates the localization information and multiscale structural information from MAB and CGM, resulting in the feature map of RAM1 being close to the ground truth of the input image. The model's discrimination ability between targets and distractors possibly comes from capturing the global

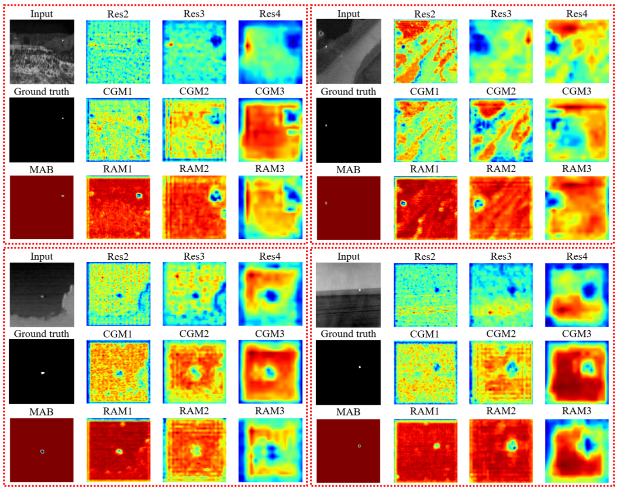

Fig. 10. Visualization of the feature maps from different modules in the proposed framework on NUST-SIRST dataset.

and local features and perceiving contextual information by the MAB and RAM.

V. Conclusion

An effective deep learning framework decoupled with priors is proposed to detect infrared small targets and mainly contains CGM, MAB, and RAM. The CGM is integrated into multilevel features to enable the framework to capture local features, surround context, and global context information. The MAB exchanges multiscale context of low level and high level to richly decode semantic information and spatial details of small targets. The RAM is designed to combine the features from MAB and CGM for aggregating the localization information and multiscale structural information. The unite loss function consisting of BCE loss and IoU loss is used to weigh the positive and negative classes. Extensive experiments demonstrate that our designed framework has higher detection accuracy and lower false alarm rate on infrared small target detection in complex backgrounds compared with some other models. The AUC, IoU, and F values of our model are evaluated to 0.9510, 0.4766, and 0.6909 on NUST-SIRST dataset, and the interference time is 12.12 s for dealing with 200 images. The metrics of AUC, IoU, and F for our model on NUDT-SIRST dataset reach 0.9534, 0.6245, and 0.7654. The proposed robust and effective data-driven method may shed light on infrared searching and tracking applications.

However, there are some limitations to the proposed method. Our framework makes it difficult to discriminate the targets within edge regions of the input image. In the future, we will develop a multiframe detection method by synthesizing spatial and temporal features of infrared small target videos.

References

[1] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, "Single-frame infrared small-target detection: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 87–119, Jun. 2022, doi: 10.1109/MGRS.2022.3145502.

[2] Z. Yang, T. Ma, Y. Ku, Q. Ma, and J. Fu, "DFFIR-net: Infrared dim small object detection network constrained by gray-level distribution model," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5026215, doi: 10.1109/TIM.2022.3220285.

[3] S. Cao, J. Deng, J. Luo, Z. Li, J. Hu, and Z. Peng, "Local convergence index-based infrared small target detection against complex scenes," *Remote Sens.*, vol. 15, no. 5, 2023, Art. no. 1464.

[4] F. Xiangsuo, Q. Wenlin, L. Juliu, H. Qingnan, and Z. Fan, "Dim and small target detection based on spatio-temporal filtering and high-order energy estimation," *IEEE Photon. J.*, vol. 15, no. 2, Apr. 2023, Art. no. 7800420, doi: 10.1109/JPHOT.2023.3242991.

[5] H. Yao, L. Liu, Y. Wei, D. Chen, and M. Tong, "Infrared small-target detection using multidirectional local difference measure weighted by entropy," *Sustainability*, vol. 15, no. 3, Jan. 2023, Art. no. 1902.

[6] H. Yi et al., "Spatial-temporal tensor ring norm regularization for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 7000205, doi: 10.1109/LGRS.2023.3236030.

[7] R. Li and Y. Shen, "YOLOSR-IST: A deep learning method for small target detection in infrared remote sensing images based on super-resolution and YOLO," *Signal Process.*, vol. 208, Jul. 2023, Art. no. 108962.

[8] Z. Chen, Q. Sheng, J. Li, B. Wang, and C. Yin, "Fast and robust infrared small target detection based on a combination of hand-designed features and machine-designed features," *Remote Sens. Lett.*, vol. 14, no. 3, pp. 221–230, Feb. 2023.

[9] R. Kou et al., "Infrared small target segmentation networks: A survey," *Pattern Recognit.*, vol. 143, Nov. 2023, Art. no. 109788.

[10] J. Lin, K. Zhang, X. Yang, X. Cheng, and C. Li, "Infrared dim and small target detection based on U-transformer," *J. Vis. Commun. Image Representation*, vol. 89, Nov. 2022, Art. no. 103684.

[11] H. Gong et al., "Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images," *Remote Sens.*, vol. 14, no. 12, Jun. 2022, Art. no. 2861.

[12] L. Deng, H. Zhu, Q. Zhou, and Y. Li, "Adaptive top-hat filter based on quantum genetic algorithm for infrared small target detection," *Multimedia Tools Appl.*, vol. 77, pp. 10539–10551, Mar. 2018.

[13] Yubo Zhang, L. Zheng, and Y. Zhang, "Small infrared target detection via a Mexican-hat distribution," *Appl. Sci.*, vol. 9, no. 24, Dec. 2019, Art. no. 5570.

[14] Z. Wang, S. Duan, and C. Sun, "Infrared small target detection method combined with bilateral filter and local entropy," *Secur. Commun. Netw.*, vol. 2021, Feb. 2021, Art. no. 6661852.

[15] 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, doi: 10.1109/TGRS.2013.2242477.

[16] 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, doi: 10.1109/JSTARS.2022.3193884.

[17] X. Cao, C. Rong, and X. Bai, "Infrared small target detection based on derivative dissimilarity measure," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3101–3116, Aug. 2019, doi: 10.1109/JSTARS.2019.2920327.

[18] 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, doi: 10.1109/TIP.2013.2281420.

[19] X. Chen, W. Xu, S. Tao, T. Gao, Q. Feng, and Y. Piao, "Total variation weighted low-rank constraint for infrared dim small target detection," *Remote Sens.*, vol. 14, no. 18, Sep. 2022, Art. no. 4615.

[20] F. Yan, G. Xu, Q. Wu, J. Wang, and Z. Li, "Infrared small target detection using kernel low-rank approximation and regularization terms for constraints," *Infrared Phys. Technol.*, vol. 125, Sep. 2022, Art. no. 104222.

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

[22] F. Yan, G. Xu, J. Wang, Q. Wu, and Z. Wang, "Infrared small target detection via schatten capped pnorm-based non-convex tensor low-rank approximation," *IEEE Trans. Geosci. Remote Sens.*, vol. 20, 2023, Art. no. 6002105, doi: 10.1109/LGRS.2022.3227550.

[23] C. Gao and Y. Zhai, "Region proposal patch-image model for infrared small target detection," *Int. J. Remote Sens.*, vol. 43, pp. 424–456, Feb. 2022.

[24] X. Li, Y. Li, R. Wu, C. Zhou, and H. Zhu, "Short circuit recognition for metal electrorefining using an improved faster R-CNN with synthetic infrared images," *Front. Neurorobot.*, vol. 15, Nov. 2021, Art. no. 751037.

[25] K.R. Akshatha, A. K. Karunakar, S. B. Shenoy, A. K. Pai, N. H. Nagaraj, and S. S. Rohatgi, "Human detection in aerial thermal images using faster R-CNN and SSD algorithms," *Electronics*, vol. 11, no. 7, Apr. 2022, Art. no. 1151.

[26] L. Sommer, A. Schumann, T. Müller, T. Schuchert, and J. Beyerer, "Flying object detection for automatic UAV recognition," in *Proc. IEEE 14th Int. Conf. Adv. Video Signal Based Surveill.*, Lecce, Italy, 2017, pp. 1–6, doi: 10.1109/AVSS.2017.8078557.

[27] L. Ding, X. Xu, Y. Cao, G. Zhai, F. Yang, and L. Qian, "Detection and tracking of infrared small target by jointly using SSD and pipeline filter," *Digit. Signal Process.*, vol. 110, Mar. 2021, Art. no. 102949.

[28] X. Mou, S. Lei, and X. Zhou, "YOLO-FR: A YOLOv5 infrared small target detection algorithm based on feature reassembly sampling method," *Sensors*, vol. 23, Mar. 2023, Art. no. 2710.

[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.*, Seoul, South Korea, 2019, pp. 8508–8517.

[30] 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.*, Waikoloa, HI, USA, 2021, pp. 949–958, doi: 10.1109/WACV48630.2021.00099.

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

[32] 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. 4250–4260, Aug. 2023, doi: 10.1109/TAES.2023.3238703.

[33] Y. Chuang et al., "Infrared small target detection based on multiscale local contrast learning networks," *Infrared Phys. Technol.*, vol. 123, Jun. 2022, Art. no. 104107.

[34] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002013, doi: 10.1109/TGRS.2022.3163410.

[35] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2023, doi: 10.1109/TIP.2022.3199107.

[36] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023, doi: 10.1109/TIP.2022.3228497.

[37] T. Wu et al., "MTU-Net: Multilevel TransUNet for space-based infrared tiny ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601015, doi: 10.1109/TGRS.2023.3235002.

[38] P. Pan, H. Wang, C. Wang, and C. Nie, "ABC: Attention with bilinear correlation for infrared small target detection," Mar. 17, 2023, *arXiv:2303.10321*. [Online]. Available: http://arxiv.org/abs/2303.10321

[39] R. Kou et al., "LW-IRSTNet: Lightweight infrared small target segmentation network and application deployment," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5621313, doi: 10.1109/TGRS.2023.3314586.

[40] H. Sun, J. Bai, F. Yang, and X. Bai, "Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5000513, doi: 10.1109/TGRS.2023.3235150.

[41] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.

[42] M. Mokari-Mahallati, R. Ebrahimpour, N. Bagheri, and H. Karimi-Rouzbahani, "Deeper neural network models better reflect how humans cope with contrast variation in object recognition," *Neurosci. Res.*, vol. 192, pp. 48–55, Jul. 2023.

[43] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3902–3911, doi: 10.1109/CVPR.2019.00403.

[44] B. Li et al., "AEFusion: A multi-scale fusion network combining axial attention and entropy feature aggregation for infrared and visible images," *Appl. Soft Comput.*, vol. 132, Jan. 2023, Art. no. 109857.

[45] K. Wang, Li Liu, X. Fu, L. Liu, and W. Peng, "RA-DENet: Reverse attention and distractions elimination network for polyp segmentation," *Comput. Biol. Med.*, vol. 155, Mar. 2023, Art. no. 106704.

[46] M. Montazerolghaem, Y. Sun, G. Sasso, and A. Haworth, "U-Net architecture for prostate segmentation: The impact of loss function on system performance," *Bioengineering*, vol. 10, no. 4, Mar. 2023, Art. no. 412.

[47] S. Zhong, H. Zhou, X. Cui, X. Cao, F. Zhang, and J. Duan, "Infrared small target detection based on local-image construction and maximum correntropy," *Measurement*, vol. 211, Apr. 2023, Art. no. 112662.

[48] J. Xing and M. Jia, "A convolutional neural network-based method for workpiece surface defect detection," *Measurement*, vol. 176, May 2021, Art. no. 109185.

[49] C. Guo, X. Chen, Y. Chen, and C. Yu, "Multi-stage attentive network for motion deblurring via binary cross-entropy loss," *Entropy*, vol. 24, Sep. 2022, Art. no. 1414.

[50] X. Nan and L. Ding, "Multi-scale attention and structural relation graph for local feature matching," *IEEE Access*, vol. 10, pp. 110603–110615, 2022, doi: 10.1109/ACCESS.2022.3215168.

[51] X. Xiao et al., "BASeg: Boundary aware semantic segmentation for autonomous driving," *Neural Netw.*, vol. 157, pp. 460–470, Jan. 2023.

[52] M. Qi et al., "FTC-Net: Fusion of transformer and CNN features for infrared small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8613–8623, 2022, doi: 10.1109/JSTARS.2022.3210707.

Shunshun Zhong is currently working toward the Ph.D. degree with the College of Mechanical and Electrical Engineering, Central South University, Changsha, China.

His research interests include computer vision and image processing.

Fan Zhang received the Ph.D. degree from the College of Mechanical and Electrical Engineering, Central South University, Changsha, China, in 2019.

He is currently an Associate Professor with the School of Automation, Central South University. His research interests include image processing, computer vision, and pattern recognition.

Ji'an Duan received the Ph.D. degree from the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1996.

He is currently a Professor with the State Key Laboratory of High Performance Complex Manufacturing, Central South University, Changsha, China. He is a Cheung Kong Scholars Chair Professor of China. His research interests include computer vision, optical device fabrication, and encapsulation.