Infrared Small and Dim Target Detection With Transformer Under Complex Backgrounds

Fangcen Liu, Chenqiang Gao, Fang Chen, Deyu Meng, *Member, IEEE*, Wangmeng Zuo, *Senior Member, IEEE*, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— The infrared small and dim (S&D) target detection is one of the key techniques in the infrared search and tracking system. Since the local regions similar to infrared S&D targets spread over the whole background, exploring the correlation amongst image features in large-range dependencies to mine the difference between the target and background is crucial for robust detection. However, existing deep learning-based methods are limited by the locality of convolutional neural networks, which impairs the ability to capture large-range dependencies. Additionally, the S&D appearance of the infrared target makes the detection model highly possible to miss detection. To this end, we propose a robust and general infrared S&D target detection method with the transformer. We adopt the self-attention mechanism of the transformer to learn the correlation of image features in a larger range. Moreover, we design a feature enhancement module to learn discriminative features of S&D targets to avoid miss-detections. After that, to avoid the loss of the target information, we adopt a decoder with the U-Net-like skip connection operation to contain more information of S&D targets. Finally, we get the detection result by a segmentation head. Extensive experiments on two public datasets show the obvious superiority of the proposed method over state-of-the-art methods, and the proposed method has a stronger generalization ability and better noise tolerance.**

***Index Terms*— Transformer, infrared small and dim target, detection.**

Manuscript received 8 November 2021; revised 6 July 2022, 28 November 2022, and 7 April 2023; accepted 10 October 2023. Date of publication 26 October 2023; date of current version 1 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62176035, Grant U22A2096, and Grant 12226004; in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJZD-K202100606; and in part by the Chongqing Graduate Research Innovation Project under Grant CYB22249. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Senem Velipasalar. *(Corresponding author: Chenqiang Gao.)*

Fangcen Liu, Chenqiang Gao, and Xinbo Gao are with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China, and also with the Chongqing Key Laboratory of Signal and Information Processing, Chongqing 400065, China (e-mail: liufc67@gmail.com; gaocq@cqupt.edu.cn; gaoxb@cqupt.edu.cn).

Fang Chen is with the School of Electrical Engineering and the Computer Science Department, University of California at Merced, Merced, CA 95343 USA (e-mail: fchen905@usc.edu).

Deyu Meng is with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China, and also with the Henan Engineering Research Center for Artificial Intelligence Theory and Algorithms, School of Mathematics and Statistics, Henan University, Kaifeng, Henan 475004, China (e-mail: dymeng@mail.xjtu.edu.cn).

Wangmeng Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 478221, China (e-mail: wmzuo@hit.edu.cn).

Digital Object Identifier 10.1109/TIP.2023.3326396

I. Introduction

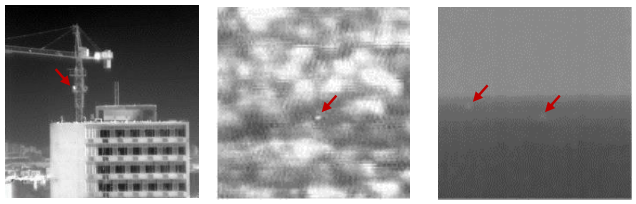
THE infrared S&D target detection is one of the key techniques in the infrared search and tracking (IRST) system because the infrared imaging can capture targets from a long distance and has a strong anti-interference ability [1], [2], [3]. However, this task encounters kinds of challenges [4], as illustrated in Fig. 1. Infrared targets in infrared images are S&D, while backgrounds are usually complex. As a result, the S&D target is easily submerged in the complex background, with a low Signal-to-Clutter Ratio (SCR). In addition, the number of target pixels is much fewer than background pixels, which leads to that the target and background pixels in an image are of extreme imbalance.

To address the above challenges, model-driven infrared S&D target detection methods routinely design the detection model for these targets by deeply mining the prior knowledge of imaging characteristics of targets, backgrounds, or both of them [5], [6], [7]. However, these approaches heavily rely on prior knowledge, which makes their generalization ability limited. Additionally, model-driven approaches can not be fast and easily applied to a new sample whose imaging characteristic does not well match the assumption of the model. In contrast, data-driven models are more feasible and can easily adapt to a new sample. In recent years, deep learning methods are adopted to detect infrared S&D targets and have shown stronger generalization ability and feasibility [4], [8].

However, among existing deep learning methods, feature learning mainly relies on convolutional neural networks (CNNs). The locality of CNNs impairs the ability to capture large-range dependencies [9], which easily results in high false alarms. As can be observed from Fig. 1, the local regions similar to infrared S&D targets spread over the whole background. Thus, it is very important to learn the difference between the target and the background in a large range.

Currently, the transformer structure, from the Natural Language Processing (NLP) field [10], [11], has demonstrated its powerful ability in non-local feature learning in various computer vision tasks [12], [13]. Different from CNN architectures, the transformer architecture contains the self-attention mechanism and the feed-forward network. The self-attention mechanism enables the transformer to have the ability to capture large-range dependencies of all embedded tokens.

In this paper, we adopt the transformer to learn the correlation amongst all embedded tokens of an image. Firstly,



(a) Building background (b) Cloudy background (c) Sea-sky background

Fig. 1. Illustration of challenges of the infrared S&D target detection task under complex backgrounds. These targets in different scenes appear pretty small, dim, and sparse, which makes the detection model easy to miss detection. From (a) to (c), we can also obviously observe that the local regions similar to infrared S&D targets spread over the whole background, which easily leads to high false alarms.

we embed an image into a sequence of tokens by the Resnet-50 [14]. After that, the self-attention mechanism is adopted to model complex dependencies among different embedded tokens, so that the difference between the S&D target and background can be well mined.

Furthermore, due to the small size and the dim appearance of the target, as can be seen from Fig. 1, if we can not capture discriminative information of S&D targets, it would highly likely lead to miss-detections. To this end, we design a feature enhancement module as the feed-forward network to acquire more discriminative features of these targets. Moreover, since S&D target features are easily lost in the network, we adopt a U-Net-like [15] upsampling structure to get more information of targets.

We summarize the main contributions of the paper as follows:

- We propose a novel S&D target detection method. It adopts the self-attention mechanism of the transformer to learn the correlation amongst all embedded tokens so that the network can learn the difference between the S&D target and background in a larger range. To our best knowledge, this is the first work to explore the transformer to detect the infrared S&D target.
- The designed feature enhancement module can help learn more discriminative features of S&D targets.
- We evaluate the proposed method on two public datasets and extensive experimental results show that the proposed method is effective and significantly outperforms state-of-the-art methods.

The remainder of this paper is organized as follows: In Section II, related works are briefly reviewed. In Section III, we present the proposed method in detail. In Section IV, the experimental results are given and discussed. Conclusions are drawn in Section V.

II. Related Work

A. Small Target Detection

1) Infrared Small-Dim Target Detection: In the early stages, the model-driven infrared S&D target detection methods design filters to enhance the target or suppress the background [16], [17]. Zeng et al. [18] and Deshpande et al. [19] proposed the Top-Hat method and max-mean/max-median method to directly enhance targets by filtering them out from original images, respectively. Then, Deng et al. [17] proposed an adaptive M-estimator ring top-hat transformation method to detect the S&D target. Aghaziyarati et al. [20] and Moradi et al. [21] detected S&D targets by suppressing the estimated background as much as possible. However, the detection performances of these methods are limited by designed filters.

Inspired by the human visual system, some efforts are based on the different local contrast which focus on the saliency of the target to distinguish the target from the background and improve the performance of infrared S&D target detection [22], [23], [24], [25]. Deng et al. [23] and Gao et al. [24] focused on the saliency of the target to distinguish the target from the background. Cui et al. [26] proposed an infrared S&D target detection algorithm with two layers which can balance those detection capabilities. The first layer was designed to select significant local information. Then the second layer leveraged a classifier to separate targets from background clutters.

Based on the observation that infrared S&D targets often present sparse features, while the background has the non-local correlation property [27], [28]. Gao et al. [7] firstly proposed the patch-image in infrared S&D target detection and the low-rank-based infrared patch-image (IPI) model. Dai et al. [29] proposed a column weighted IPI model (WIPI), and then proposed a reweighed infrared patch-tensor model (RIPT) [30]. However, the model-driven approaches heavily rely on prior knowledge, which makes the generalization ability of these models limited.

Recently, the generalization ability of the data-driven infrared S&D target detection is well promoted by deep learning methods [31], [32], [33], [34], [35], [36], [37], [38], [39]. Fan et al. [40] designed a convolutional neural network architecture to improve the contrast between the S&D target and the background. Zhao et al. [41] proposed a TBC-Net which included a target extraction module and semantic constraint module. Wang et al. [4] used adversarial generation networks (GANs) to balance Miss Detection (MD) and False Alarm (FA). Shi et al. [37] and Zhao et al. [31] regarded S&D targets as noises, and they treated the S&D targets detection task as a denoising task. Hou et al. [38] combined handcrafted feature methods and convolutional neural networks and proposed a robust RISTDnet framework. Dai et al. [8], [42] combined the global and local information of the infrared image and proposed end-to-end detection methods named ALC and ACM to solve the problem of the lack of fixed features of infrared S&D targets.

Compared with the above data-driven methods, the proposed method overcomes the shortcoming of the limitation of the CNN to learn the correlation amongst all embedded tokens, which can mine differences between targets and backgrounds in a larger range.

2) Small Target Detection in RGB Images: Different from infrared S&D target methods, small target detection methods for RGB images payed more attention to solving the problem of small target size [43], [44], and usually adopted data augmentation [45], multi-scale learning [46] and context information learning [47] strategies to improve the robustness and generalization of detection. However, using the above

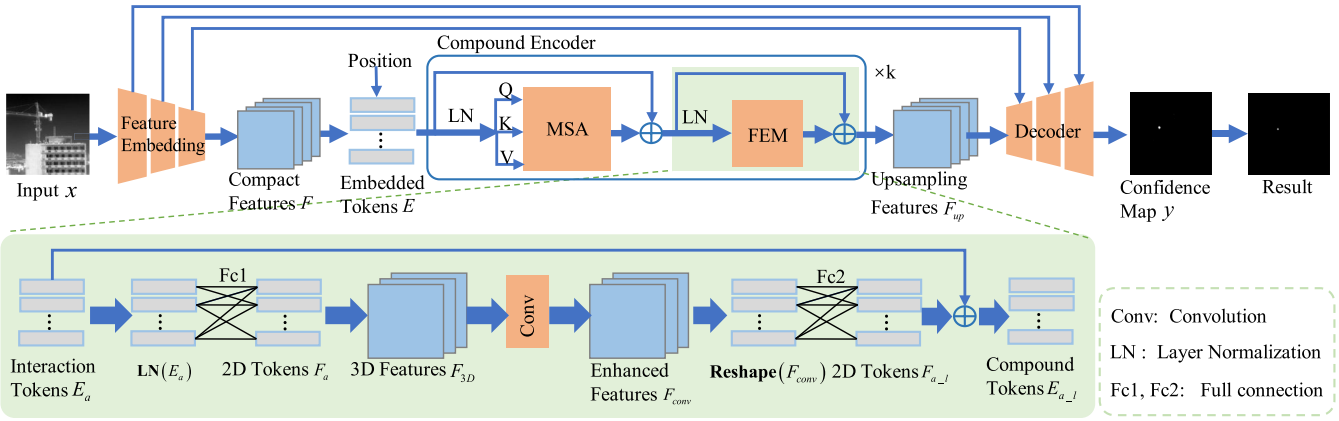


Fig. 2. The proposed infrared S&D target detection framework in this study. It contains three parts: a feature embedding module, a compound encoder with k encoder layers, and a decoder. The feature embedding module is proposed to obtain the compact feature representation. In the compound encoder, each encoder layer mainly has two parts: the multi-head self-attention (MSA) mechanism and the feature enhancement module (FEM). The multi-head self-attention is adopted to learn the correlation amongst all embedded tokens. The designed feature enhancement module can learn more discriminative features of S&D targets. The decoder is adopted to obtain the confidence map of the S&D target.

methods directly to detect infrared S&D targets would make the performance drop sharply, as verified in [4]. Compared with targets in RGB images, infrared S&D targets have high similarity to the background with a low SCR, so it is more difficult to distinguish S&D targets from backgrounds. Moreover, the max-pooling layer adopted in these methods may suppress or even eliminate features of infrared S&D targets [48].

B. Combining Self-Attention Mechanisms With CNNs

Inspired by the success of the self-attention mechanism adopted in transformer architectures in the NLP field [10], [11], some works employed them in the computer vision field [49], [50], [51].

Vision Transformer (ViT) is the pioneering work that directly applied a transformer architecture on non-overlapping medium-sized image patches for image classification [13]. Liu et al. [52] proposed a hierarchical Transformer structure named swin transformer. Such hierarchical architecture had the flexibility to model at various scales and had linear computational complexity with respect to image size. The most obvious advantage of the transformer is its ability to capture the image information through the self-attention mechanism in a large range. However, the self-attention mechanism's performance in local information learning is relatively limited compared with CNN-based methods. Hence, some methods proposed to combine the strengths of CNNs and self-attention mechanisms [53], [54], [55]. Carion et al. [56] adopted Resnet-50 or Resnet-101 [14] to acquire the compact feature representation, and then introduced this representation into the self-attention mechanism. Liu et al. [52] proposed a general-purpose transformer backbone for computer vision. D'Ascoli et al. [9] proposed the gated positional self-attention to mimic the locality of convolutional layers. Chen et al. [57] combined self-attention mechanism with U-net [15] for medical segmentation.

Different from these methods, the proposed method in this study focuses on the small size of the target, and designs a feature enhancement module to learn more discriminative features of S&D targets.

III. Proposed Method

A. Overview

As depicted in Fig. 2, the proposed method consists of three main modules: (1) A feature embedding module to extract a compact feature representation of an image. (2) A compound encoder to learn correlation amongst all embedded tokens and more discriminative features of S&D targets. (3) A decoder to produce confidence maps.

Given an image of size $C \times H \times W$, we embed it into a sequence of tokens by the Resnet-50 [14]. Then the designed compound encoder is used to learn the correlation amongst all embedded tokens and capture more discriminative features of S&D targets. After that, with the help of the U-Net-like [15] skip connection operation, embedded features are concatenated with feature maps obtained by the decoder to obtain the confidence map. Finally, we segment the confidence map to obtain the detection result.

B. CNN-Based Feature Embedding Module

In ViT, an image is divided into a sequence of non-overlapping patches of the same size and then use the linear mapping to embed these patches to a sequence of tokens [13]. In this study, the proposed method adopts the Resnet-50 [14] as the feature embedding module to extract compact features of the original image, and then reshapes them into a sequence of tokens.

After the input image $x \in \mathbb{R}^{C\times H\times W}$ passes through the feature embedding module, compact features $F \in \mathbb{R}^{C_1\times H'\times W'}$ with local information are obtained. Then we flatten 3D features F into 2D tokens $E_{em} \in \mathbb{R}^{H'W'\times C_2}$, where $H'W'$ is the number of tokens. We learn specific position embeddings E_{pos} to maintain the spatial information of these features. Finally, we obtain embedded tokens $E = E_{em} + E_{pos}$, where $E \in \mathbb{R}^{n\times C_2}$ and $E = (E_1, E_2, \cdots, E_n)$, n is the number of tokens, and $n = H'W'$.

C. Compound Encoder

The compound encoder contains k encoder layers, and all encoder layers have the same structure. An encoder layer includes a multi-head self-attention module with m heads and a feature enhancement module. The multi-head self-attention mechanism aims to capture the correlation amongst n tokens so that differences between targets and backgrounds can be well constructed. The feature enhancement module aims to learn more discriminative features of S&D targets.

1) Multi-Head Self-Attention Module: Embedded tokens E are divided into m heads $E = \left\{E^1, E^2, \cdots, E^m\right\}$, $E^j \in \mathbb{R}^{n\times\frac{C_2}{m}}$, and then fed into the multi-head self-attention module MSA$(\cdot)$ to obtain interaction tokens E_a, we define these processes as:

$$E_a = \mathrm{MSA}\left(\mathrm{LN}\left(E\right)\right) + E, \tag{1}$$

where the LN$(\cdot)$ is the layer normalization.

In each head, the multi-head self-attention module MSA$(\cdot)$ defines three learnable weight matrices to transform Queries ($W^Q \in \mathbb{R}^{C_2\times\frac{C_2}{m}}$), Keys ($W^K \in \mathbb{R}^{C_2\times\frac{C_2}{m}}$) and Values ($W^V \in \mathbb{R}^{C_2\times\frac{C_2}{m}}$). The embedded tokens E^j of a head are first projected onto these weight matrices to get $Q^j = EW_j^Q$, $K^j = EW_j^K$ and $V^j = EW_j^V$. The output $Z^j \in \mathbb{R}^{n\times\frac{C_2}{m}}$ of the self-attention layer is given by:

$$Z^j = \mathrm{softmax}\left(\frac{Q^j {K^j}^T}{\sqrt{\frac{C_2}{m}}}\right) V^j, \tag{2}$$

where j is the j-th head of the multi-head self-attention. The result of m heads can be expressed as:

$$Z = \mathrm{Concate}\left(Z^1, Z^2, \cdots, Z^m\right) W^z, Z \in \mathbb{R}^{n\times C_2} \tag{3}$$

where the projections are parameter matrix $W^z \in \mathbb{R}^{C_2\times C_2}$.

2) Feature Enhancement Module: We feed interaction tokens E_a into the designed feature enhancement module to obtain compound tokens E_{a_l}.

Specifically, the feature enhancement module is shown in Fig. 2. Firstly, these interaction tokens E_a are fed into the first full connection layer to obtain 2D tokens $F_a = \left(F_{a\_1}, F_{a\_2}, \cdots, F_{a\_n}\right)$, $F_a \in \mathbb{R}^{n\times C_3}$. Then we reshape 2D tokens into 3D features F_{3D} with the size of $n \times P \times P$ and adopt the convolution operation to learn the local information of F_{3D}, which helps enhance the features of S&D targets. Finally, enhanced features F_{conv} are further reshaped back to the size of $n \times C_3$ and then fed into the next full connection layer to learn the next 2D tokens F_{a_l}. After that, with the summation of F_{a_l} and E_a, we obtain compound tokens $E_{a\_l} \in \mathbb{R}^{n\times C_2}$.

D. Feature Decoder With Skip Connection

We adopt a decoder to upsample reshaped compound tokens $F_{up} \in \mathbb{R}^{C_2\times H'\times W'}$ to obtain confidence maps of S&D targets. All features in the feature embedding process are concatenated with feature maps obtained by upsampling operation through skip connection operation like U-Net structure [15] to prevent the loss of the S&D target contextual information. After that, the confidence map y is obtained.

TABLE I

THE RESOLUTIONS OF DIFFERENT DATASTES

Resolutions	Training Sets		Testing Sets	
	Number (MFIRST)	Nubmer (SIRST)	Number (MFIRST)	Nubmer (SIRST)
(0, 200) × (0, 200]	9960	2	47	1
[138, 200) × (200, 400]	0	95	16	31
[200, 400) × (200, 410]	0	241	36	52
[278, 400] × [418, 592]	0	3	1	2

TABLE II

THE RATIO OF SMALL TARGETS TO THE WHOLE IMAGE. R_{MAX}, R_{MIN}, AND R_{MEAN} MEAN THE MAXIMUM, MINIMUM, AND AVERAGE VALUE OF THE RATIO OF SMALL TARGETS TO THE WHOLE IMAGE, RESPECTIVELY

	Training Sets		Testing Sets	
	MFIRST	SIRST	MFIRST	SIRST
R_{max}	1.07%	0.32%	0.56%	0.48%
R_{min}	0.031%	0.005%	0.003%	0.006%
R_{mean}	0.24%	0.06%	0.078%	0.06%

E. Loss Function

To handle the class imbalance issue between targets and backgrounds [8] and focus more on S&D target regions, the Intersection of Union (IoU) loss is adopted to calculate the distance between the confidence map and the ground truth, defined by:

$$L_{iou} = 1 - \frac{y \cap x_{gt}}{y \cup x_{gt}}. \tag{4}$$

The y is the confidence map, and the x_{gt} is the ground truth image.

IV. EXPERIMENT

A. Experimental Setup

1) Datasets: We adopt the widely used MFIRST dataset [4] and SIRST dataset [42] to evaluate the proposed method. The MFIRST dataset contains 9960 training samples and 100 test samples. Among them, all infrared S&D target image samples are generated by the random combination of real backgrounds and real S&D targets or simulated targets with Gaussian spatial gray distribution. The SIRST dataset is a widely-used public dataset that contains 341 training samples and 86 test samples.

In these datasets, S&D targets usually appear in the sea, sky, mountains, or buildings background. However, according to our statistics, two datasets differ in the image resolution and the ratio of small targets to the whole image, and the statistical results are listed in Tables I and II. As shown in these tables, the resolution of the SIRST dataset is larger than the MFIRST dataset, and the ratio of small targets to the whole image in the training set and the testing set are consistent on the SIRST dataset but inconsistent on the MFIRST dataset. These differences lead to a higher detection difficulty on the MFIRST dataset.

2) Evaluation Metrics: As the same as [7], we regard that the detection is correct when the following two conditions are met simultaneously: (1) The output result has some overlap pixels with the ground truth. (2) The pixel distance between

TABLE III
FEATURE DIMENSIONS OF THE INPUTS AND OUTPUTS OF EACH MODULE

Modules	Operation		Input size	Output size
Feature embedding module	Root		3×224×224	64×112×112
	Convolution		64× 112×112	64×55×55
	Body 0		64×55×55	256×55×55
	Body 1		256×55×55	512×28×28
	Body 2		512×28×28	1024×14×14
	Convolution		1024×14×14	768×14×14
	Flatten		768×14×14	768×196
	Transpose		768×196	196×768
	Position_embedding		196×768	196×768
Compound encoder	MSA		196×768	196×768
	FEM	Fc1	196×768	196×3136
		Reshape	196×3136	196×56×56
		Convolution	196×56×56	196×56×56
		Flatten	196×56×56	196×3136
		Fc2	196×3136	196×768
Decoder	Convolution		768×14×14	512×14×14
	Block 0		512×14×14	256×28×28
	Block 1		256×28×28	128×56×56
	Block 2		128×56×56	64×112×112
	Block 3		64×112×112	16×224×224
	Confidence map		16×224×224	2×224×224

TABLE IV
THE KEY PARAMETER SETTINGS OF ALL METHODS

Methods	Key parameter settings	Methods	Key parameter settings
Top-Hat [18]	structure size: 12×12	MPCM [58]	scale size: 3, 5, 7
Max-Mean& Max-Median [19]	filter size: 15 ×15	TLLCM [22]	gaussian kernel size: 3×3; scale size: 3, 5, 7, 9
AAGD [20]	scale size: 3, 5, 7, 9	LEF [59]	$\alpha = 0.5$, $h = 0.2$ scale size: 3, 5, 7, 9
ADMD [21]	scale size: 3, 5, 7, 9	GST [24]	$\sigma_1 = 0.6$, $\sigma_2 = 1.1$ boundary width: 5; filter size: 5×5
LIG [60]	k=0.2, N=11	ACM [42]	image size: 480×480; lr: 0.05; epoch: 300; batch size: 8; backbone: fpn; fuse: asymbi
IPI [7]	patch size: 50×50; sliding step: 10; $\epsilon = 10^{-6}$	MDvsFA [4]	image size: 128×128; lr: $10^{-}4$, $10^{-}5$; epoch: 30; batch size: 10; $\lambda_1 = 100$, $\lambda_2 = 10$
ILCM [61]	subblock size: 8×8; moving step: 4	DNANet [62]	image size: 256×256 lr: 0.05; epoch: 1500 batch size: 16; backbone: Resnet34
LPNetGA [63]	image size: 120×120; lr: 0.001; epoch: 600; batch size: 8; patch size: 30×30; sliding step: 10	Ours	image size: 224×224; lr: 0.01; epoch: 150; batch size: 24; patch head of MSA: 12; number of encoder layer: 12

the centers of the detection result and the ground truth is less than a threshold (4 pixels).

In this study, five widely used evaluation metrics [4], [7], including the probability of detection (P_d), the area under $P_d - F_a$ curve (AUC), target-level F_1 measure (F_1^t), pixel-level F_1 measure (F_1^p) and the intersection over union (IoU), are used for performance evaluation. These metrics are defined as follows:

$$P_d = \frac{\#\text{ number of true detections}}{\#\text{ number of real targets}}, \tag{5}$$

$$F_a = \frac{\#\text{number of false detections}}{\#\text{ number of images}}. \tag{6}$$

As the same as [7], we adopt P_d with F_a=0.2 and the area under $P_d - F_a$ curve (AUC) with $F_a < 0.5$ to evaluate average performance of the proposed method. The P_d can measure the correct detection rate of the model when $F_a = 0.2$. It is worth noting that we adopt the interpolation method to calculate a reasonable value as P_d based on the P_d values around $F_a = 0.2$ when P_d is missing at $F_a = 0.2$. The AUC can reflect the false alarm rate of the model. A model with a low false alarm rate has a higher AUC value.

$$F_1^t = \frac{2 \times Precision_t \times Recall_t}{Precision_t + Recall_t}, \tag{7}$$

$$F_1^p = \frac{2 \times Precision_p \times Recall_p}{Precision_p + Recall_p}, \tag{8}$$

where the precision and recall can be defined as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

where TP is the true positive, FP is the false positive and FN is the false negative. It is worth noting that the associated TP, FP, and FN represent the number of pixels when we use the pixel-level metric F_1^p. The associated TP, FP, and FN represent the number of targets when we use the target-level metric F_1^t.

The IoU is a pixel-level metric, which is defined as follows:

$$IoU = \frac{\#\text{Area of Overlap}}{\#\text{Area of Union}}. \tag{11}$$

3) Implementation Details: The framework of the proposed method is implemented using Pytorch 1.7.1, and accelerated by CUDA 11.2. The whole network is trained with the SGD algorithm with a learning rate of 0.01, momentum 0.9, and weight decay 1e-4 on NVIDIA GeForce RTX 3090 GPU. The batch size is set to 24 and the maximum epochs are set to 150. Before training, all input images are normalized. After that, these images are processed by random image flip for data augmentation. Finally, these images are resized to 224 × 224. In terms of the results of the hyperparameter discussion in the ablation study, the head of the MSA is set to 12, and the number of the encoder layer is set to 12. The feature dimensions of the inputs and outputs of each module are in Table III to show the details of the whole framework and the key parameter settings of all methods are listed in Table IV.

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

We compare the proposed method with following related methods:

- Model-driven methods: Top-Hat [18], Max-Mean/Max-Median [19], AAGD [20], ADMD [21], LIG [60], IPI [7], ILCM [61], MPCM [58], TLLCM [22], LEF [59], GST [24].
- Deep learning methods: MDvsFA [4], ACM [42], DNANet [62] and LPNetGA [63].

1) Quantitative Evaluation: The results of different methods on two public datasets [4], [42] are listed in Table V. From this table, we can observe that deep-learning-based methods are significantly superior to model-driven methods. The proposed method outperforms both model-driven methods and deep-learning-based methods.

On the MFIRST dataset, the proposed method outperforms the MDvsFA method by 3.46% on P_d, 7.56% on AUC, 7.32% on F_1^t and 4.23% on F_1^p. Because the MDvsFA method aims to balance outputs of two detection generators (There are two generators G_1 and G_2. G_1 aims to minimize MD, G_2 aims to

TABLE V

COMPARISON ON DIFFERENT DATASETS. '−' MEANS THAT THE METHOD CAN NOT GET REASONABLE VALUES UNDER FIXED $F_a = 0.2$ FOR P_d OR UNDER $F_a \leqslant 0.5$ FOR AUC

Methods	MFIRST					SIRST					Times
	$P_d(\%)$	AUC(%)	$F_1^t(\%)$	$F_1^p(\%)$	IoU(%)	$P_d(\%)$	AUC(%)	$F_1^t(\%)$	$F_1^p(\%)$	IoU(%)	(ms/image)
Top-Hat [18]	-	-	44.62	12.80	6.84	85.34	82.38	82.52	44.13	28.33	9.8
Max-Mean/Max-Media [19]	-	50.50	58.30	14.44	7.78	78.46	77.45	73.49	23.97	13.63	8.4
AAGD [20]	43.66	56.96	65.70	32.42	19.35	89.09	88.14	84.69	50.27	33.41	37.0
ADMD [21]	59.64	64.09	70.99	31.52	18.71	94.13	90.46	88.50	56.69	39.37	16.2
LIG [60]	59.29	64.17	70.87	41.27	26.00	90.19	90.00	89.72	59.15	41.91	718.6
IPI [7]	41.59	51.02	60.73	33.58	20.18	86.87	84.45	85.32	56.97	39.79	4243.0
ILCM [61]	-	-	24.52	0.91	-	-	-	47.26	0.71	-	14.5
MPCM [58]	57.86	64.62	72.20	35.43	21.53	93.56	90.40	86.96	58.59	41.38	51.1
TLLCM [22]	-	46.43	52.63	6.67	3.44	61.61	79.14	79.66	7.60	3.89	767.1
LEF [59]	49.49	70.01	72.45	5.87	3.02	-	-	59.6	2.45	1.23	5549.2
GST [24]	56.39	59.69	66.67	24.67	14.70	77.01	76.81	80.40	35.32	21.46	7.6
MDvsFA [4]	86.62	81.78	85.27	60.36	42.89	-	-	-	-	-	14.6
ACM [42]	70.07	71.95	82.11	58.05	40.89	98.24	91.67	96.78	81.30	68.50	**7.4**
DNANet [62]	76.81	76.45	84.71	60.72	43.59	97.20	97.22	97.70	**84.19**	**72.70**	45.7
LPNetGA [63]	85.71	84.27	88.11	59.36	42.21	97.23	94.81	95.93	70.54	54.52	883.7
Ours	**90.08**	**89.34**	**92.59**	**64.59**	**47.70**	**100.00**	**99.14**	**98.62**	83.16	71.18	50.9

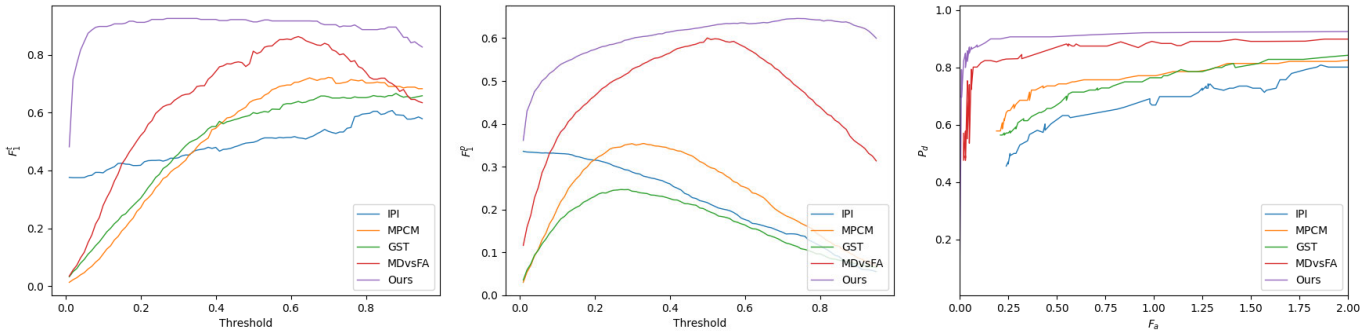


Fig. 3. The F_1^t and F_1^p curves with different thresholds and the $P_d - F_a$ ROC curve on MFIRST dataset.

minimize FA), but does not focus on exploring the difference between the target and the background. Compared with the ACM, DNANet and LPNetGA methods, the proposed method outperforms a lot. Because the proposed method not only learns the correlation amongst all embedded tokens, but also extracts more discriminative features of S&D targets from a local view.

On the SIRST dataset, the proposed method outperforms the ACM and LPNetGA methods in all metrics, while outperforming DNANet in target-level metrics (P_d, AUC, F_1^t). According to our statistical analysis in Tables I and II, the SIRST dataset has a larger image resolution, and consistent distribution of the small target sizes compared to the MFIRST dataset. Thus, with the help of the dense nested interactive skip connection module, the DNANet method can achieve better results in pixel-level metrics. However, in a more difficult dataset, the proposed method has stronger robustness and stability.

The F_1^t, F_1^p and $P_d - F_a$ receiver operating characteristic (ROC) curves of some representative methods on MFIRST dataset are shown in Fig. 3. As we can see from the figure, the proposed method has better performance than other methods at different segmentation thresholds. These results validate the robustness of the proposed method.

In addition, we calculate the FLOPs of the proposed method and the deep-learning-based methods. As can be seen in Tables V, the FLOPs of MDvsFA [4], ACM [42], DNANet [62], LPNetGA [63], and the proposed method are 61.78G, 1.01G, 19.99G, 22.04G, and 42.65G, respectively.

To evaluate the inference time of the proposed method and the state-of-the-art methods, we use a machine equipped with an Intel Xeon Silver 4210 @ 2.20GHz CPU and an NVIDIA RTX 3090 GPU, and results are presented in Tables V. We can see from the table that the inference time of the proposed method is reasonable. It should be noted that there is a slight discrepancy between the reported inference times in this study and those reported in the original papers for the comparison methods, due to differences in experimental settings and running conditions. The experimental settings of this study can be seen in Tables IV.

2) Qualitative Evaluation: Fig. 4 shows qualitative evaluation comparisons of five representative methods. It can be seen that all targets in infrared images with different complex backgrounds appear small, dim, and sparse, but the proposed method can achieve the best detection performance. Most methods, including ADMA, MDvsFA, and ACM have some false alarms when detecting the S&D target with a high noise level, a low SCR, and with building shelters. The MDvsFA method fails to detect the right target when detecting the S&D target with bright clutters. Some methods are difficult to acquire robust detection when the infrared target is pretty dim and the contrast between the background and the target is extremely weak. This can be seen in the

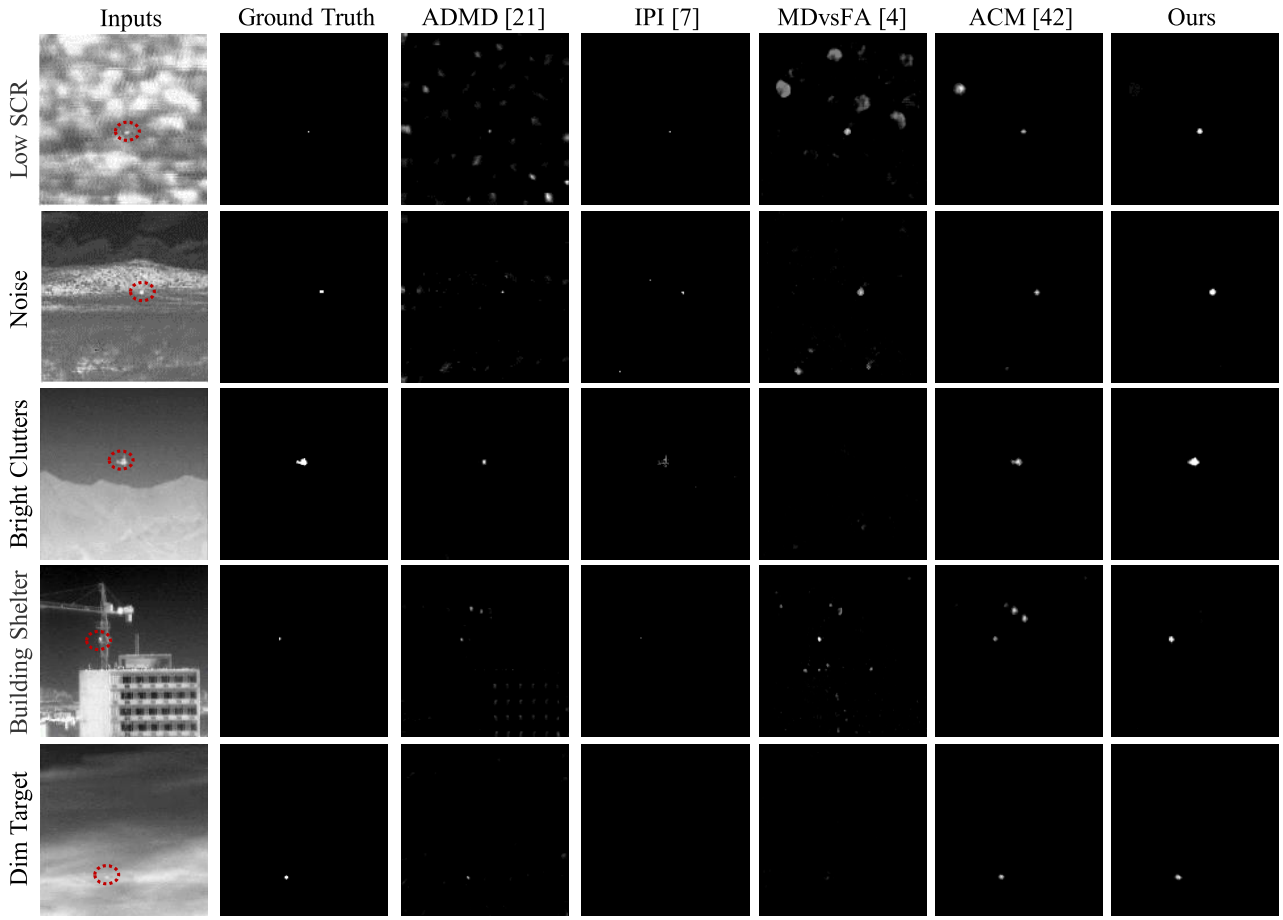


Fig. 4. The representative processed the results of different methods. The first row to the last row shows the results of detecting an S&D target with a low SCR, a high noise level, bright clutters, complex building shelters, and a dim appearance, respectively.

last row of Fig. 4 where the S&D target is hidden in the clouds.

The proposed method can learn the correlation amongst embedded tokens in a larger range, and focus on learning more discriminative features of S&D targets, so it has the best detection results in all kinds of complex backgrounds.

C. Ablation Study

1) Effeteness of Different Modules: In this section, we validate the effeteness of the compound encoder which helps to learn the correlation amongst all embedded tokens and more discriminative features of S&D targets. The results are listed in Table VI. The 'Baseline' method [57] for comparison consists of three modules: a Resnet-50 structure [14], a self-attention encoder that has the same structure as the ViT [13] model, a decoder with skip connection. The 'Baseline w/o MSA' method removes the self-attention encoder from the baseline structure. The 'Baseline w/o pooling' method removes the pooling layer from the Resnet-50 structure. The 'Baseline w/o pooling w/FEM' method is the proposed method.

Experimental results in Table VI obviously show the promotion effect of each component on infrared S&D target detection. As can be seen, the baseline method with the normal transformer encoder can achieve reasonable results. All metrics degrade when we remove the self-attention encoder.

TABLE VI
THE PERFORMANCE OF THE MODULE ABLATION STUDIES ON DIFFERENT DATASETS

Methods	MFIRST		SIRST	
	P_d(%)	AUC(%)	P_d(%)	AUC(%)
Baseline	80.00	80.76	99.78	97.98
Baseline w/o MSA	79.28	80.39	97.66	97.30
Baseline w/o pooling	87.05	84.77	99.90	99.00
Baseline w/o pooling w/ FEM (Ours)	**90.08**	**89.34**	**100.00**	**99.14**

All metrics improve a lot by removing the pooling layer. These results show that the pooling layer can degrade the feature learning ability of S&D targets, which is the same as [48]. The performances are further improved after adopting a feature enhancement module (FEM) in the compound encoder. On MFIRST dataset, the P_d is improved by 3.03%, and the AUC is improved by 4.57%. On SIRST dataset, the P_d is improved by 0.10%, and the AUC is improved by 0.14%.

Further, we visualize the feature heatmaps in the last layer of the decoder module after adding and removing the FEM, respectively, to investigate the effectiveness of the FEM on infrared small targets, as shown in Fig. 5. From the figure, we can see that the model has a stronger learning ability for infrared small targets and effectively reduces the cases of miss-detections when we add the FEM in the compound

TABLE VII

THE PERFORMANCE OF THE PROPOSED METHOD AND THE ACM METHOD IN THE SIMILAR NUMBER OF PARAMETERS.'*' MEANS WE MODIFIED THE METHODS.'XX_448', 'XX_224', AND 'XX_480' MEANS THE INPUT SIZE ARE 448 × 448, 224 × 224, AND 480 × 480, RESPECTIVELY

Methods	MFIRST				SIRST				Parameters
	$P_d(\%)$	AUC(%)	$F_1^t(\%)$	$F_1^p(\%)$	$P_d(\%)$	AUC(%)	$F_1^t(\%)$	$F_1^p(\%)$	
ACM_480	70.07	71.95	82.11	58.05	98.24	91.67	96.78	81.30	0.39
Ours_224	**90.08**	**89.34**	**92.59**	**64.59**	**100.00**	**99.14**	**98.62**	**83.16**	110.64
ACM*_224	78.89	70.63	87.60	64.31	98.67	92.93	95.85	78.62	93.22
Ours*_224	**88.39**	**86.28**	**91.97**	**65.05**	**100.00**	**98.38**	**97.30**	**80.97**	80.52
ACM*_480	72.14	66.42	65.58	54.78	99.88	89.83	98.17	87.42	93.22
Our*_448	**86.96**	**84.93**	**88.41**	**62.96**	**100.00**	**99.77**	**98.63**	**88.07**	89.94

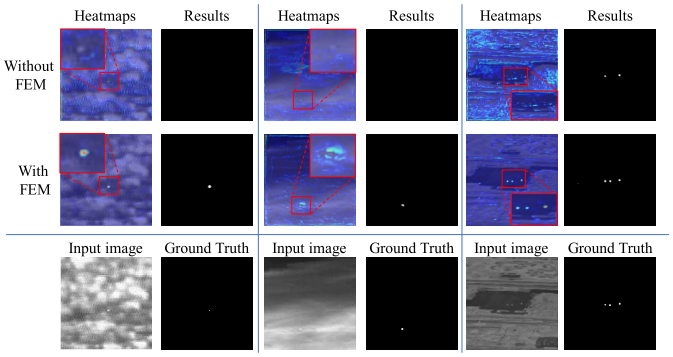


Fig. 5. The feature heatmaps of the last layer of the decoder module after adding and removing the FEM.

encoder. It shows that the local information learned by the feature enhancement module helps learn more discriminative features of S&D targets.

2) Model Size Discussion: We adjust the ACM method and the proposed method to be of close model sizes, and then evaluate the performance to verify that the performance improvement of the proposed method is not entirely due to the increase in the number of parameters. Specific adjustments are as follows:

- ACM: Change the number of network channels from [8, 16, 32, 64] to [64, 128, 512, 1024].
- The proposed method: Reduce the number of encoder layers in the compound encoder module from 12 to 8.

The experimental results are listed in Table VII. As we can see from this table, increasing the number of model parameters could lead to performance improvement. Aiming at the S&D characteristics of the infrared target, the proposed method designs a feature enhancement module and applies a self-attention mechanism to the infrared small target detection task. The designed feature enhancement module can help to learn more discriminative features of S&D targets and the adopted self-attention mechanism of the transformer can learn the correlation amongst all embedded tokens so that the proposed model can learn the difference between the small target and background in a larger range. These designs help to improve detection performance, so the performance improvement of the proposed method is not entirely due to the increase in the number of parameters.

3) Hyperparameter Discussion: The number of encoder layers and the head number of the MSA module influence the effect of learning the correlation amongst all embedded tokens. In this section, we investigate these important hyper-parameters, the results can be seen in Table VIII.

TABLE VIII

THE PERFORMANCE OF THE HYPERPARAMETER ABLATION STUDIES ON THE MFIRST DATASET

Encoder layers	Head	$P_d(\%)$	AUC(%)
14	3	84.39	84.38
14	12	85.71	85.08
12	3	87.14	83.11
12	12	**90.08**	**89.34**
6	3	85.00	84.26
6	12	85.71	85.52

As can be observed in Table VIII, the dependency between the target and background could not be constructed adequately as the number of encoder layers decreases. The head number of MSA has little effect on performance. Finally, based on the best experimental performance, we set the number of the encoder layers to 12, and the head number of the MSA module to 12.

D. Performance Comparison of Different Feature Embedding Modules

The feature embedding module is important to extract sufficient information for the compound encoder module to capture long-range dependencies between tokens, and learn S&D target discriminative features. Hence, we explore the performance of the proposed method by combining other feature embedded modules. The experimental results are listed in Table IX, in which 'Ours_Resnet101' and 'Ours_Resnet34' mean we adopt the Resnet-101 and the Resnet-34 as the feature embedding modules, respectively. The 'Ours_Resnet50' is the proposed method. As can be seen from the table, the model achieves the best experimental results on most metrics on both public datasets when adopting the Resnet-50. Thus, we adopt the Resnet-50 as the feature embedding module in this study.

E. Analysis on Generalization

In practice, the well-trained model is likely to be applied to a new data, so the generalization is very important. Consequently, we evaluate the generalization of the proposed model on different datasets. The experimental results are listed in Table X.

As can be seen from Table X, these experiments verify the strong generalization of deep learning methods. However, compared with other deep learning methods, the proposed

TABLE IX
THE RESULTS OF DIFFERENT FEATURE EMBEDDING MODULES

Methods	MFIRST				SIRST			
	$P_d(\%)$	AUC(%)	$F_1^t(\%)$	$F_1^p(\%)$	$P_d(\%)$	AUC(%)	$F_1^t(\%)$	$F_1^p(\%)$
Ours_Resnet34	86.33	86.10	89.47	64.61	98.15	97.45	95.93	82.46
Ours_Resnet101	87.08	86.79	92.02	**65.54**	96.30	94.22	94.34	77.97
Ours_Resnet50	**90.08**	**89.34**	**92.59**	64.59	**100.00**	**99.14**	**98.62**	**83.16**

TABLE X
THE PERFORMANCE OF DIFFERENT METHODS IN TERMS OF CROSS-DATASET GENERALIZATION. '−' MEANS THAT THE METHOD CAN NOT GET REASONABLE VALUES

Methods	MFIRST(Train) SIRST(Test)		SIRST(Train) MFIRST(Test)	
	$P_d(\%)$	AUC(%)	$P_d(\%)$	AUC(%)
MDvsFA [4]	97.22	93.95	-	-
ACM [42]	92.13	57.49	67.25	62.05
ours	**98.15**	**96.02**	**87.77**	**85.58**

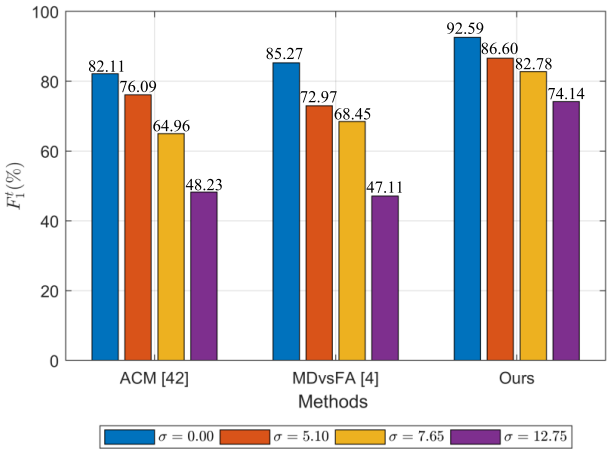


Fig. 6. The F_1^t performance of different methods with different Gaussian noises. 'σ' means the standard deviation of Gaussian noise.

method achieves better performance in terms of generalization. The proposed method outperforms the MDvsFA method by 0.93% on P_d and 2.07% on AUC when the model is trained on the MFIRST dataset and tested on the SIRST dataset. The proposed method outperforms the ACM method by 20.52% on P_d and 23.53% on AUC when the model is trained on the SIRST dataset and tested on the MFIRST dataset. Although the distribution of the dataset is heterogeneous, the differences between infrared S&D targets and background in different datasets are similar. The self-attention mechanism helps capture the correlation amongst embedded tokens to learn the difference between the target and background from a large region, so it has the best detection results under the cross-dataset situation.

Moreover, the performance of training on the training set of the MFIRST dataset and testing on the testing set of the SIRST dataset is better than that of training on the training set of the SIRST dataset and testing on the testing set of the MFIRST dataset. This is because, according to our statistics in Tables I and II, the MFIRST test dataset has a higher detection difficulty, and the number of images in the MFIRST training dataset is larger than that in the SIRST dataset.

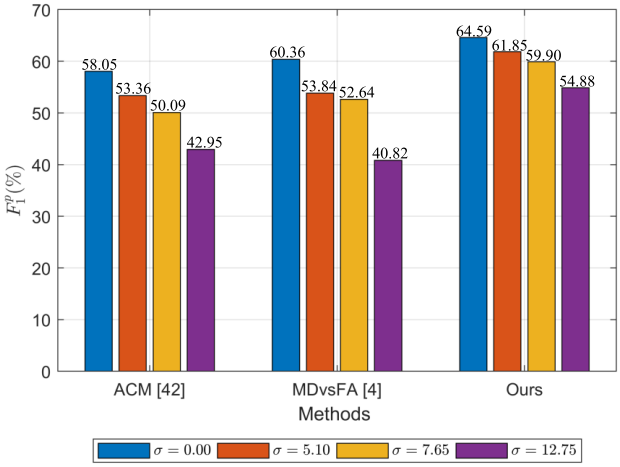


Fig. 7. The F_1^p performance of different methods with different Gaussian noises. 'σ' means the standard deviation of Gaussian noise.

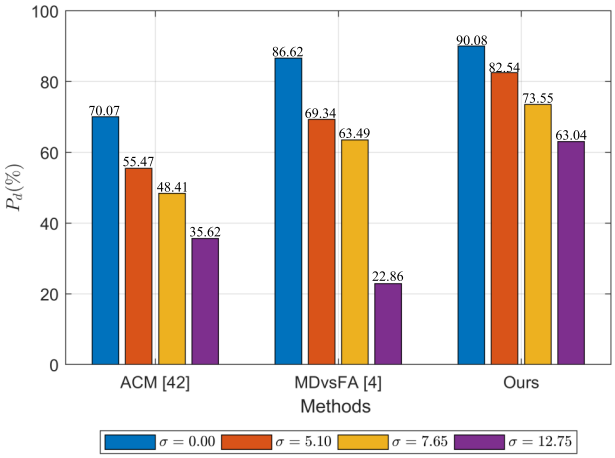


Fig. 8. The P_d performance of different methods with different Gaussian noises. 'σ' means the standard deviation of Gaussian noise.

F. Analysis on Noise Tolerance

In addition to the generalization, the robustness to noise is also very crucial. Thus, we evaluate the noise tolerance performance of the proposed model on the MFIRST dataset. We set four kinds of Gaussian noise with different standard deviations, which are 5.10, 7.65, 12.75, and 25.50, respectively. The mean of these Gaussian noises is set to 0. It's worth noting that we only add the above noise to the test set.

When the standard deviations are 0.00, 5.10, 7.65, and 12.75, the experimental results are showed in Fig. 6, 7, 8, and 9. As we can see from these figures, the detection performance gradually decreases as the noise

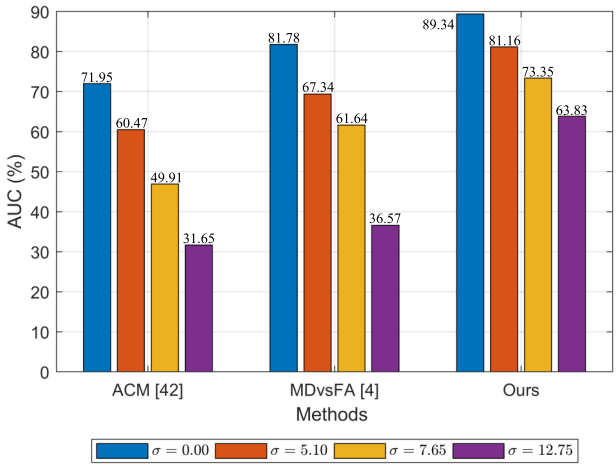


Fig. 9. The AUC performance of different methods with different Gaussian noises. 'σ' means the standard deviation of Gaussian noise.

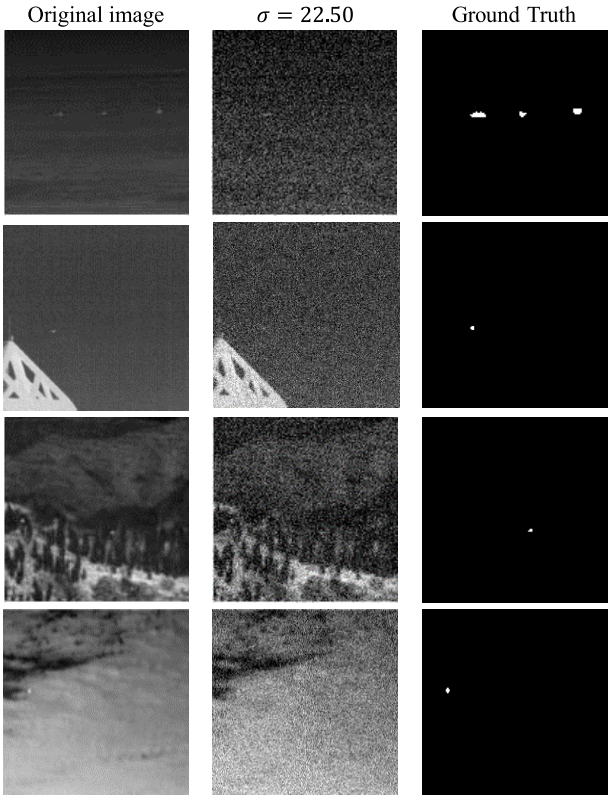


Fig. 10. From left to right, the first column is the original image, the second column is the image after adding Gaussian noise, and the third column is the ground truth. 'σ' means the standard deviation of Gaussian noise.

standard deviation increases. However, compared with other deep learning methods, the proposed method has the best performance under Gaussian noise with different standard deviations. Meanwhile, the performance of the method decreases slowly relative to other methods as the standard deviation of the noise increases. The experimental results show that the proposed method has stronger resistance to noise.

When the standard deviation is 25.50, the Fig. 10 illustrates some samples of inputs. As we can see from this figure, the S&D targets are completely submerged in backgrounds, and we can hardly find out S&D targets. The Table XI shows the detection performance of different methods, and we can see from the table that the performance of all methods deteriorates dramatically. Even so, the proposed method still has the best experimental performance. Fig. 11 shows the detection performance of different methods when the standard deviation of Gaussian noise is 25.50. From the figure we can observe that the ACM method tends to miss detection when the S&D target is not completely submerged in the background, while the MDvsFA network tends to increase false alarms. In contrast, the proposed method can robustly detect S&D targets.

TABLE XI

THE DETECTION PERFORMANCE OF DIFFERENT METHODS WHEN THE STANDARD DEVIATION OF THE GAUSSIAN NOISE IS 25.50

Methods	$F_1^t(\%)$	$F_1^p(\%)$	P_d	AUC
ACM [42]	22.76	22.65	13.90	13.85
MDvsFA [4]	24.05	23.04	18.81	18.37
Ours	**47.40**	**41.67**	**34.88**	**35.67**

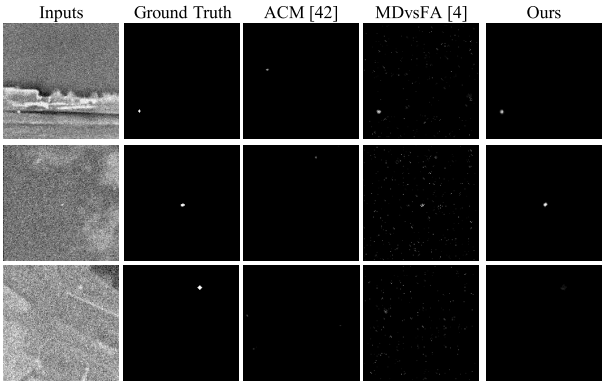


Fig. 11. The detection performance when the standard deviation of the Gaussian noise is 25.50.

V. CONCLUSION

In this paper, we propose a new infrared S&D target detection framework. We adopt the multi-head self-attention module to explore the correlation amongst all embedded tokens, and thus differences between targets and backgrounds can be well learned. In addition, the feature enhancement module is designed to learn more discriminative features of S&D targets. Experiments on two public datasets show that compared with state-of-the-art methods, the proposed method performs much better on detecting infrared S&D targets with complex backgrounds. Additionally, experimental results also show the proposed method has a stronger generalization ability and better noise tolerance.

REFERENCES

[1] H. Zhu, J. Zhang, G. Xu, and L. Deng, "Balanced ring top-hat transformation for infrared small-target detection with guided filter kernel," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 5, pp. 3892–3903, Oct. 2020.

[2] T. Zhang, Z. Peng, H. Wu, Y. He, C. Li, and C. Yang, "Infrared small target detection via self-regularized weighted sparse model," *Neurocomputing*, vol. 420, pp. 124–148, Jan. 2021.

[3] 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.

[4] 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.

[5] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Infrared small-target detection using multiscale gray difference weighted image entropy," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 1, pp. 60–72, Feb. 2016.

[6] C. Gao, L. Wang, Y. Xiao, Q. Zhao, and D. Meng, "Infrared small-dim target detection based on Markov random field guided noise modeling," *Pattern Recognit.*, vol. 76, pp. 463–475, Apr. 2018.

[7] 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.

[8] 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.

[9] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving vision transformers with soft convolutional inductive biases," 2021, *arXiv:2103.10697*.

[10] H. Ye et al., "Contrastive triple extraction with generative transformer," in *Proc. AAAI*, vol. 35, 2021, pp. 14257–14265.

[11] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[12] H. D. Nguyen, X.-S. Vu, and D.-T. Le, "Modular graph transformer networks for multi-label image classification," in *Proc. AAAI*, 2021, vol. 35, no. 10, pp. 9092–9100.

[13] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–22.

[14] 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.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[16] J. Han, C. Liu, Y. Liu, Z. Luo, X. Zhang, and Q. Niu, "Infrared small target detection utilizing the enhanced closest-mean background estimation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 645–662, 2021.

[17] L. Deng, J. Zhang, G. Xu, and H. Zhu, "Infrared small target detection via adaptive M-estimator ring top-hat transformation," *Pattern Recognit.*, vol. 112, pp. 107729–107737, Apr. 2021.

[18] M. Zeng, J. Li, and Z. Peng, "The design of top-hat morphological filter and application to infrared target detection," *Infr. Phys. Technol.*, vol. 48, no. 1, pp. 67–76, Apr. 2006.

[19] R. Venkateswarlu, "Max-mean and max-median filters for detection of small targets," *Proc. SPIE*, vol. 3809, Jul. 1999, pp. 74–83.

[20] S. Aghaziyarati, S. Moradi, and H. Talebi, "Small infrared target detection using absolute average difference weighted by cumulative directional derivatives," *Infr. Phys. Technol.*, vol. 101, pp. 78–87, Sep. 2019.

[21] S. Moradi, P. Moallem, and M. F. Sabahi, "Fast and robust small infrared target detection using absolute directional mean difference algorithm," *Signal Process.*, vol. 177, Dec. 2020, Art. no. 107727.

[22] 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.

[23] 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.

[24] C.-Q. Gao, J.-W. Tian, and P. Wang, "Generalised-structure-tensor-based infrared small target detection," *Electron. Lett.*, vol. 44, no. 23, pp. 1349–1351, 2008.

[25] J. Han, K. Liang, B. Zhou, X. Zhu, J. Zhao, and L. Zhao, "Infrared small target detection utilizing the multiscale relative local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 612–616, Apr. 2018.

[26] Z. Cui, J. Yang, S. Jiang, and J. Li, "An infrared small target detection algorithm based on high-speed local contrast method," *Infr. Phys. Technol.*, vol. 76, pp. 474–481, May 2016.

[27] D. Pang, T. Shan, W. Li, P. Ma, R. Tao, and Y. Ma, "Facet derivative-based multidirectional edge awareness and spatial–temporal tensor model for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5001015.

[28] Y. Sun, J. Yang, and W. An, "Infrared dim and small target detection via multiple subspace learning and spatial–temporal patch-tensor model," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3737–3752, May 2021.

[29] Y. Dai, Y. Wu, and Y. Song, "Infrared small target and background separation via column-wise weighted robust principal component analysis," *Infr. Phys. Technol.*, vol. 77, pp. 421–430, Jul. 2016.

[30] 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.

[31] B. Zhao, C. Wang, Q. Fu, and Z. Han, "A novel pattern for infrared small target detection with generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4481–4492, May 2021.

[32] M. Ju, J. Luo, G. Liu, and H. Luo, "ISTDet: An efficient end-to-end neural network for infrared small target detection," *Infr. Phys. Technol.*, vol. 114, May 2021, Art. no. 103659.

[33] J. Ryu and S. Kim, "Heterogeneous gray-temperature fusion-based deep learning architecture for far infrared small target detection," *J. Sensors*, vol. 2019, pp. 1–15, Aug. 2019.

[34] Z. Gao, J. Dai, and C. Xie, "Dim and small target detection based on feature mapping neural networks," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 206–216, Jul. 2019.

[35] J. Du, H. Lu, M. Hu, L. Zhang, and X. Shen, "CNN-based infrared dim small target detection algorithm using target-oriented shallow-deep features and effective small anchor," *IET Image Process.*, vol. 15, no. 1, pp. 1–15, Jan. 2021.

[36] X. Tong, B. Sun, J. Wei, Z. Zuo, and S. Su, "EAAU-Net: Enhanced asymmetric attention u-net for infrared small target detection," *Remote. Sens.*, vol. 13, no. 16, pp. 3200–3219, 2021.

[37] M. Shi and H. Wang, "Infrared dim and small target detection based on denoising autoencoder network," *Mobile Netw. Appl.*, vol. 25, no. 4, pp. 1469–1483, Aug. 2020.

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

[39] T. Zhang, S. Cao, T. Pu, and Z. Peng, "AGPCNet: Attention-guided pyramid context networks for infrared small target detection," 2021, *arXiv:2111.03580*.

[40] Z. Fan, D. Bi, L. Xiong, S. Ma, L. He, and W. Ding, "Dim infrared image enhancement based on convolutional neural network," *Neurocomputing*, vol. 272, pp. 396–404, Jan. 2018.

[41] M. Zhao, L. Cheng, X. Yang, P. Feng, L. Liu, and N. Wu, "TBC-Net: A real-time detector for infrared small target detection using semantic constraint," 2019, *arXiv:2001.05852*.

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

[43] M. Uzair, R. S. Brinkworth, and A. Finn, "Bio-inspired video enhancement for small moving target detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1232–1244, 2021.

[44] S. Zhou et al., "Hierarchical and interactive refinement network for edge-preserving salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1–14, 2021.

[45] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 566–583.

[46] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection—SNIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.

[47] A. Torralba and P. Sinha, "Statistical context priming for object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Jul. 2001, pp. 763–770.

[48] L. Liangkui, W. Shaoyou, and T. Zhongxing, "Using deep learning to detect small targets in infrared oversampling images," *J. Syst. Eng. Electron.*, vol. 29, no. 5, pp. 947–952, Oct. 2018.

[49] Y. Luo et al., "Dual-level collaborative transformer for image captioning," in *Proc. AAAI*, vol. 35, no. 3, pp. 2286–2293, May 2021.

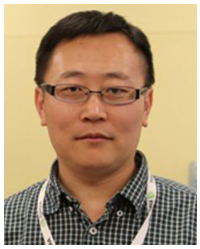
[50] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, and M. Wu, "Two-stream convolution augmented transformer for human activity recognition," in *Proc. AAAI*, vol. 35, 2021, pp. 286–293.

[51] W. Zhang, Y. Ying, P. Lu, and H. Zha, "Learning long-and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption," in *Proc. AAAI*, vol. 34, 2020, pp. 9571–9578.
[52] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[53] J. Liang, N. Homayounfar, W.-C. Ma, Y. Xiong, R. Hu, and R. Urtasun, "PolyTransform: Deep polygon transformer for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9128–9137.
[54] Z. Dai, B. Cai, Y. Lin, and J. Chen, "UP-DETR: Unsupervised pre-training for object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1601–1610.
[55] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–16.
[56] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.
[57] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[58] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, Oct. 2016.
[59] C. Xia, X. Li, L. Zhao, and R. Shu, "Infrared small target detection based on multiscale local contrast measure using local energy factor," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 157–161, Jan. 2020.
[60] H. Zhang, L. Zhang, D. Yuan, and H. Chen, "Infrared small target detection based on local intensity and gradient properties," *Infr. Phys. Technol.*, vol. 89, pp. 88–96, Mar. 2018.
[61] 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.
[62] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2023.
[63] F. Chen et al., "Local patch network with global attention for infrared small target detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 5, pp. 3979–3991, Oct. 2022.

Fangcen Liu received the B.S. and M.S. degrees in information and communication engineering from the Chongqing University of Posts and Telecommunications, China, in 2018 and 2021, respectively, where she is currently pursuing the Ph.D. degree. Her research interests include image processing, deep learning, cross-modal retrieval, and infrared small target detection.

Chenqiang Gao received the B.S. degree in computer science from the China University of Geosciences, Wuhan, China, in 2004, and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, in 2009. In August 2009, he joined the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China. In September 2012, he joined the Informedia Group, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, where he was a Visiting Scholar on multimedia event detection (MED) and surveillance event detection (SED). In April 2013, he became a Postdoctoral Fellow and continued work on MED and SED until March 2014, when he returned to CQUPT. He is currently a Professor at CQUPT. His research interests include image processing, infrared target detection, action recognition, and event detection.

Fang Chen received the B.S. degree in digital media technology from the Chongqing University of Posts and Telecommunications, China, in 2020, and the M.S. degree from the University of Southern California, Merced, USA, in 2023, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include image processing, manipulation detection, and infrared small target detection.



Deyu Meng (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2001, 2004, and 2008, respectively. He is currently a Professor with the Institute for Information and System Sciences, Xi'an Jiaotong University. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2012 to 2014. His current research interests include self-paced learning, noise modeling, and tensor sparsity.

Wangmeng Zuo (Senior Member, IEEE) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, China, in 2007. From 2004 to 2006, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University. From 2009 to 2010, he was a Visiting Professor with Microsoft Research Asia. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. He has published over 80 papers in top-tier academic journals and conferences. His current research interests include image enhancement and restoration, image generation and editing, visual tracking, object detection, and image classification. He has served as a Tutorial Organizer in ECCV 2016, an Associate Editor for the *IET Biometrics*, and a Guest Editor for *Neurocomputing*, *Pattern Recognition*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Postdoctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. He was with the School of Electronic Engineering, Xidian University, from 2001 to 2020. Since 2020, he has been the President of the Chongqing University of Posts and Telecommunications, Chongqing, China. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications. He has published five books and approximately 200 technical articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, the *International Journal of Computer Vision*, and *Pattern Recognition* in the above areas. He is a fellow of the Institution of Engineering and Technology. He has served as the general chair/co-chair, the program committee chair/co-chair, or a PC member for approximately 30 major international conferences. He is on the editorial boards of several journals, including *Signal Processing* and *Neurocomputing*.