IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 January 2024, accepted 27 February 2024, date of publication 4 March 2024, date of current version 21 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3373436

RESEARCH ARTICLE

Aware Distribute and Sparse Network for Infrared Small Target Detection

YANSONG SONG, BOXIAO WANG, AND KEYAN DONG
School of Electro-Optical Engineering, Changchun University of Science and Technology, Changchun 130000, China
Institute of Space Ophotoelectronics Technology, Changchun University of Science and Technology, Changchun 130000, China

Corresponding author: Boxiao Wang (wangboxiao1998@163.com)

This work was supported in part by the Major Science and Technology Special Projects in Jilin Province under Grant 20230301001GX and Grant 20230301002GX, and in part by the National Key Research and Development Program of China under Grant 2021YFA0718804.

ABSTRACT Deep learning has achieved tremendous success in the field of object detection. The efficient detection of infrared small targets using deep learning methods remains a challenging task. Infrared small targets are often detected in high-resolution features. Extracting high-level semantic features layer by layer in the network may lead to the loss of deep-layer targets. However, performing global detection on high-resolution feature maps results in high computational costs. To address this issue, we propose the aware distribute and sparse network (ADSNet) to preserve deep-layer small target features while accelerating inference speed. Specifically, we design the aware fusion distribute module (AFD) to aggregate global features and enhance the representation capability of deep-layer features. Subsequently, the aware cascaded sparse module (ACS) is utilized to guide step-by-step high-resolution feature sparsification. Experimental results demonstrate that the proposed method achieves accurate segmentation in various detection scenarios and for diverse target morphologies, effectively suppressing false alarms while controlling computational expenses. Ablation experiments further validate the effectiveness of each component.

INDEX TERMS Object detection, infrared imaging, infrared small target detection, feature fusion.

I. INTRODUCTION

Infrared sensors capture the radiant heat emitted by targets, allowing imaging in low-light conditions or harsh atmospheric environments, making infrared target detection an essential part of rapid warning systems [1], precise guidance [2], and ground surveillance. The increasing complexity of the battlefield environment has led to advances in defense systems, which should prioritize the accurate detection of targets as early and as far as possible. Long-range targets occupy only a few pixels in the infrared image, lack distinct shape characteristics, and are easily confused with complex low-altitude backgrounds. As such, the detection of infrared small targets presents a challenging task.

With the enhanced maneuverability of the aircraft, the large trajectory changes in image sequences significantly affect the detection performance of methods based on multi-frame [3]. Consequently, single-frame based detection methods have gained widespread attention [4]. Most traditional single-frame detection approaches depend on local contrast information between the target and the surrounding environment for infrared small target detection, specifically gray-scale discontinuity outliers in a slowly transitioning background. The Local Contrast Method (LCM) [5]interprets the position of the infrared target as the central point with the maximum contrast in the neighborhood grayscale values. The Local Energy Factor (LEF) [6], building upon LCM, introduces local dissimilarity to enrich the description of local differences. The Tri-layer Local Contrast Measure (TLLCM) [7] proposes core, reserve, and surrounding layers with filtered windows, each purposefully enhancing the contrast between the core and the background to extract multi-scale targets. These traditional methods are based on manual assumptions about infrared small target characteristics. Such assumptions cannot accurately distinguish background clutter that resembles preset features and are susceptible to false alarms [8]. Traditional methods are also highly sensitive to hyper-parameters like preset window

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

size and segmentation threshold, requiring specific tuning for different scenarios. When dealing with non-point targets that have a complex structure, it is challenging to accurately predict the segmentation boundary, resulting in poor generalization.

Supervised deep learning methods are driven by labeled data and modify the feature parameters according to the set loss function, which can classify multiple labels and localize to achieve target detection, and play an important role in various fields such as transportation, military, and people's livelihoods [9]. In the realm of infrared small target detection, several deep learning-based approaches have been proposed. Attentional Local Contrast Networks (ALCNet) [4] and Dense Nested Attention Network (DNANet) [10] emphasize the contextual information for small targets and high-level feature fusion. ALCNet employs bottom-up attention modulation to highlight and preserve small target features, incorporating a cyclic shift accelerating scheme for long-distance information contextual interaction. DNANet introduces dense nested interaction module and channel and spatial attention module to achieve progressive feature fusion and adaptive feature enhancement. In addition, Interior Attention-Aware Network (IAANet) [11] proposed a two-stage segmentation method, utilizing the Region Proposal Network (RPN) structure to obtain target bounding boxes for feature extraction, followed by encoding the feature map with the Transformer structure to acquire aware features. ALCNet and DNANet enhance the deep feature representation of small targets, but produce more redundant computations for global segmentation of small targets, while IAANet introduces additional computations for target potential region prediction as well.

In this paper, inspired by the preservation of deep features for small targets through multiscale feature fusion and the use of sparse patches to accelerate inference in remote sensing images, we propose the Aware Distribute and Sparse Network (ADSNet) for efficient infrared small target detection. ADS consists of two key modules. Firstly, we design the Aware Fusion Distribute module (AFD) to integrate global features, enhance attention aware, and overcome the loss of details caused by passing information layer by layer in the feature map. Subsequently, to accelerate inference, we design Aware Cascaded Sparse module (ACS), activating multi-instance perception and enriching the criteria for decision mask judgment. Afterward, deep-layer features predict decision masks layer by layer, guiding the sparsification of shallow features to enhance detection efficiency. The main contributions of this article are summarized as follows.

1) We propose an ADSNet that utilizes instance activation to sparse shallow features to efficient object detection for small infrared targets.

2) An Aware Fusion Distribute module is proposed to enrich the deep features to avoid the disappearance of small target information.

3) We conducted extensive analysis on the NUDT-SIRST [10] and IRSTD-1k [12] datasets. Experimental results demonstrate that our approach outperforms existing state-of-the-art algorithms for small infrared target detection while maintaining low computational costs.

II. RELATED WORKS

A. INSTANCE SEGMENTATION

Instance segmentation requires algorithms to assign a pixel-level mask with category labels for each instance in the image [13]. Two-stage instance segmentation methods consist of a bounding box detection stage and a mask segmentation stage, such as Mask-RCNN methods [14], [15]. One-stage methods like RetinaNet [16] and CondInst [17] eliminate the region proposals and directly generate mask segmentation for improved detection efficiency. However, it may not be conducive to identifying small targets. To enhance the accuracy of segmentation masks, some scholars focus on instance edges. Contour-based [18], [19] methods attempt to generate a coarse initialization of the contour and then iteratively regress each edge node to obtain the final instance contour. Boundary Refinement [20] methods re-predict the coarse mask to recover the lost details during upsampling. These methods perform well in conventional object segmentation but cannot be directly applied to infrared small targets.

B. SMALL OBJECT DETECTION

When dealing with the challenges of few pixels, limited features, and imbalanced target background samples that arise with deep learning for detecting small targets, researchers suggest four potential solutions: 1) multi-scale feature fusion [21], [22], 2) adding contextual information [23], [24], [25], 3) balancing class examples [24], and 4) super-resolution [24]. In addition, some researchers concentrate computational resources on the potential regions of the target to improve detection efficiency. QueryDet [27] predicts target positions at low resolution and guides high-resolution sparse predictions. OAN [28] divides remote sensing images into patches, uses a lightweight fully convolutional network to determine if a patch contains a target, and applies the detection head only to patches with objects. We aim to use various implicit activation methods to provide a more stable decision mask, cascade sparse feature layers to improve detection efficiency.

C. MULTI-SCALE FEATURE FUSION

In order to realize the interaction of target detail information in high-resolution feature maps with high-level semantic information in low-resolution feature maps. The efficient cross-scale fusion of features by FPNs has pioneered this type of research. EfficientDet [29] repeatedly utilizes BiFPN for bidirectional fusion of information. YoloF [30] employs dilated encoder and uniform matching to achieve single-level feature map detection. In the context of small infrared target

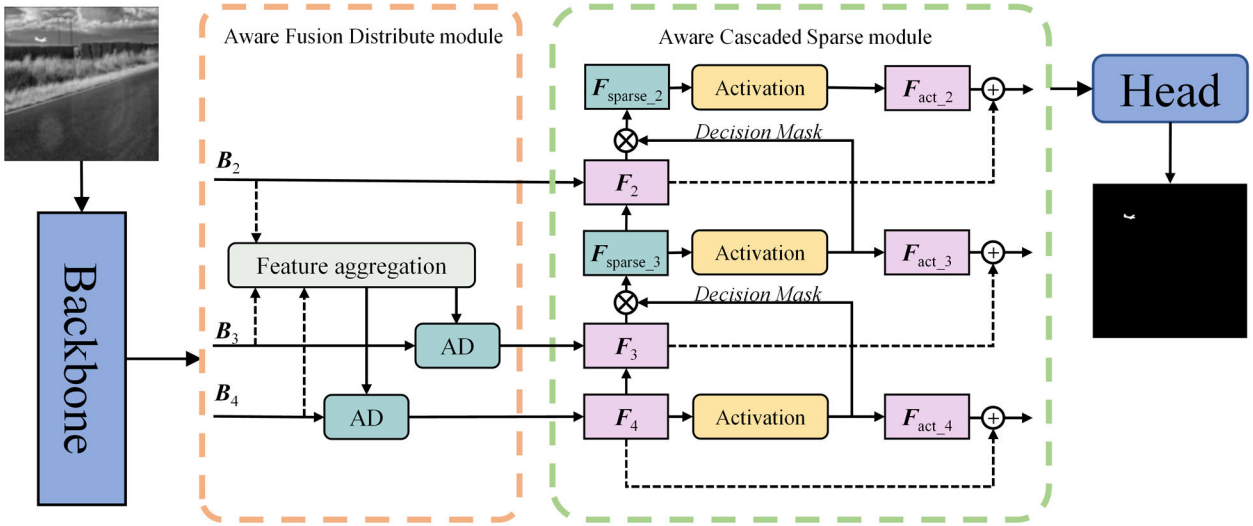

FIGURE 1. The architecture of the proposed ADSNet.

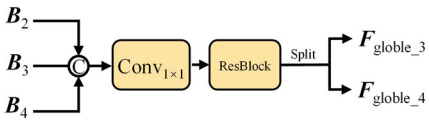

FIGURE 2. The process of feature aggregation.

detection, ALCNet [4] proposes a bottom-up feature fusion to retain detailed features. However, small infrared target features are limited, and the layer-wise transmission leads to more information loss. To address this issue, we focus on highlighting the features of weak small targets and implementing cross-layer fusion.

D. SPARSE SIGNAL PROCESSING

Signal sparsifiability is applied in many fields such as radar signal processing, image processing, information clustering, etc., while sparse information processing techniques have been developed tremendously in the past decades, and only a few representative works are presented here. In order to enhance the ISAR image recovery performance, [31] introduced a multiple measurement vector sparse recovery model to obtain sparser and more accurate results. Reference [32] proposed a re-weighted total generalized variation model to denoise the image while preserving the target edges. Reference [33] uses a paradigm to efficiently generate sparse solutions to enhance stable multiview clustering in complex noise for high dimensional data.

III. METHODOLOGY

In this section, we provide a detailed description of the proposed end-to-end infrared small target detection network. The overall architecture of ADSNet is illustrated in Fig. 1. From the base object detection network, we employ a multiscale deep feature pyramid. The Aware Fusion Distribute module integrates multi-scale features, preserving information for small targets. Then, the Aware Cascaded Sparse module utilizes activation maps to predict sparse scores for decision mask prediction, implementing a cascaded sparsity from coarse to fine, thus decreasing the computational cost of high-resolution feature maps.

A. AWARE FUSION DISTRIBUTE MODULE

During the process of layer-wise upsampling, small infrared target features are relatively fragile, making them susceptible to information loss due to operations such as pooling. Additionally, the layer-wise propagation of features exacerbates the loss of information. The intricate details of small targets in high-resolution feature maps are challenging to transmit effectively to deeper layers. In order to break this information transfer limitation and realize feature cross-layer information interaction, we propose Aware Fusion Distribute module. It aggregates and fuses information from different layers to form global information. This global information is then injected into different levels, reinforcing the preservation of information for small targets in the FPN. as illustrated in Fig. 1. Intermediate features $\boldsymbol{B}_l$ ($l = 2,3,4$) in the down-sampling stage of the backbone network are taken as input, and the downsampling step size is set to 2^l.

The feature aggregation process is shown in Fig. 2. We align the features to maintain computational efficiency while preserving small target information. we choose $\boldsymbol{B}_3$ as the feature alignment size consideration criterion. Specifically, we use max pooling and bilinear interpolation to adjust the sizes of the $\boldsymbol{B}_2$ and $\boldsymbol{B}_4$ feature maps to match the size of $\boldsymbol{B}_3$. The feature maps are concatenated, and 1×1 convolution is employed to adjust the channel sizes. Then, a residual block is utilized to further enhance the fused feature map, and the output is $\boldsymbol{F}_{\text{global}}$. The feature map generated by the residual

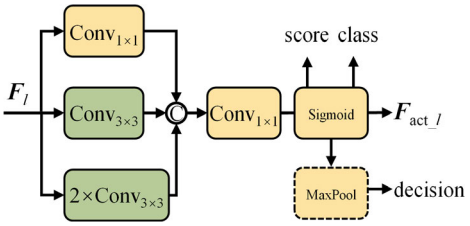

FIGURE 3. The process of feature activation. The green-highlighted portion indicating the replacement of regular convolutions with sparse convolutions during the 2nd and 3rd stages of inference.

block is split into two groups, namely $\boldsymbol{F}_{\text{global\_3}}$ and $\boldsymbol{F}_{\text{global\_4}}$, for the fusion of features at different levels. The formula is as follows:

$$\boldsymbol{F}_{\text{global}} = Res(concat[\mathcal{F}_{\text{align}}(\boldsymbol{B}_2), \boldsymbol{B}_3, \mathcal{F}_{\text{align}}(\boldsymbol{B}_4)]) \quad (1)$$

where $\mathcal{F}_{\text{align}}$ denotes the alignment of $\boldsymbol{B}_l$ to the size of $\boldsymbol{B}_3$.

To effectively distribute global information to different levels, an attention-enhanced mechanism is employed to fuse information. The aware distribute process is illustrated in Fig. 3. After pixel-wise addition of $\boldsymbol{B}_l$ and $\boldsymbol{F}_{\text{global\_}l}$, the combined features undergo channel attention (CA) branch and spatial attention (SA) branch to generate a 1D channel attention map $\boldsymbol{F}_c \in \mathbb{R}^{C_i \times 1 \times 1}$ and a 2D spatial attention map $\boldsymbol{F}_s \in \mathbb{R}^{1 \times H_i \times W_i}$, respectively. CA captures the channel responses in each spatial location and aggregates the information scattered across the channels. We perform channel dimension max-pooling on the feature maps, compressing them into a one-dimensional tensor, so that the inverse gradient computation process only targets the maximum values in the feature maps. SA simulates the localization process of the human eye to quickly target visually salient locations, and we compute the maximum values in the feature maps to ensure that the model focuses on the target's potential locations. The global feature distribution process can be summarized as follows:

$$\boldsymbol{F}' = \mathcal{F}_{\text{align}}(\boldsymbol{B}_l) + \boldsymbol{F}_{\text{global\_}l} \quad (2)$$

$$\boldsymbol{F}_{\text{c}} = \sigma(\mathcal{B}(\mathcal{C}_2^{1\times1}(\delta(\mathcal{B}(\mathcal{C}_1^{1\times1}(\mathcal{P}_{\text{max}}^{c}(\boldsymbol{F}')))))) \otimes \boldsymbol{F}_{\text{global\_}l} \quad (3)$$

$$\boldsymbol{F}_{\text{s}} = \sigma(\mathcal{B}(\mathcal{C}_3^{1\times1}(\mathcal{P}_{\text{max}}^{\text{s}}(\boldsymbol{F}')))) \otimes \boldsymbol{F}_{\text{global\_}l} \quad (4)$$

$$\boldsymbol{F} = (\boldsymbol{F}_{\text{c}} + \boldsymbol{F}_{\text{s}}) \otimes \boldsymbol{F}_{\text{local}} + \boldsymbol{F}_{\text{global\_}l} \quad (5)$$

where $\mathcal{F}_{\text{align}}$ denotes the alignment of $\boldsymbol{B}_l$ to the size of $\boldsymbol{F}_{\text{global}}$, $\mathcal{P}_{\text{max}}^{c}$ denotes max-pooling, $\mathcal{P}_{\text{max}}^{\text{s}}$ denotes channel-wise max-pooling, $\mathcal{C}^{1\times1}$ represents a 1×1 convolution, $\mathcal{B}$ denotes batch normalization, δ denotes the ReLu function, and σ denotes the sigmoid function.

B. AWARE CASCADED SPARSE MODULE

Small targets are often detected in high-resolution feature maps. However, computing the entire image for sparsely distributed targets can lead to a waste of computational resources. To address this issue, we utilize an activation map to perform cascaded sparsification on the high-resolution feature map. Instance activation maps are instance-aware

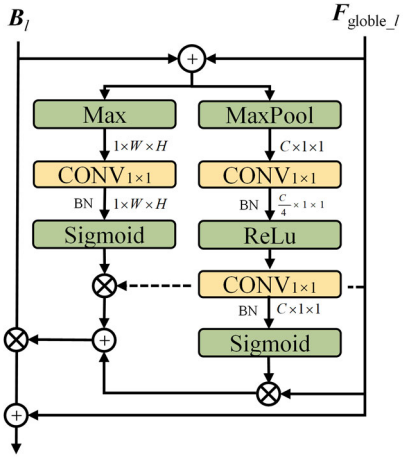

FIGURE 4. The process of aware distribute.

weighted maps that utilize implicit activation methods (e.g., categorization [34] or segmentation [35]) to highlight objects, which can uncover richer semantic information about each target. In this paper, we guide the sparsification of shallower levels by predicting decision masks at deeper levels. Although deep-level features may not be suitable for detecting smaller targets, they can still infer their approximate locations and provide high-level semantic information, guiding the subsequent stages of small target detection, as illustrated in Fig. 1.

Dense Activation takes $\boldsymbol{F}_4$ as input. In the deep-level network, target features gradually become blurred. To recover different feature details, we use a multi-scale convolutional structure. Meanwhile, we employ multi-angle supervision to activate instance awareness. Specifically, Dense Activation employs three encoding structures in parallel, 1×1, 3×3, and two sets of 3×3 convolutions, to achieve different receptive fields. The results of the three sets are concatenated, and 1×1 convolution is used to reduce the feature dimension. Subsequently, three linear layers are used for classification, decision mask, and IoU-aware score. The specific structure is shown in Fig. 4.

Decision Mask. In order to achieve dynamic sparsification of the feature map, we plan to divide the input image into $N = H_M W_M$ patches (and discuss the impact of the value of Non the detection results in Table 7). Then, we predict a binary decision mask $\boldsymbol{M} \in \{0, 1\}^N$ to determine whether to retain each patch. All elements in the decision mask are initialized to 1 and gradually updated. We additionally apply max pooling before predicting the decision mask for the $\boldsymbol{F}_3$ layer to ensure that the resolution of the decision mask prediction matches that of the $\boldsymbol{F}_4$ layer. To address the issue of the non-differentiability of the binary decision mask, making end-to-end training impractical, we use the Gumbel-Softmax [36] trick to transform the predictive probabilities p_{m} into mask:

$$\boldsymbol{M} = \text{Gumbel-Softmax}(p_{\text{m}}) \quad (6)$$

TABLE 1. Parameter settings of traditional methods.

Method	Hyperparameters Settings
Top-hat	Nhood=ones(5)
LEF	h=0.2, α=0.5, P=9
AADCDD	internal window size={3, 5, 7, 9}, external window size=19
TLLCM	window size={3, 5, 7, 9}, k=9

IoU-aware score. In the prediction of small targets, calculating the IoU leads to a highly imbalanced foreground-background ratio. Matching the ground truth to each target one-to-one can result in excessive redundant predictions for the background, and it is prone to misaligning the classification scores with the segmented targets. The paper [37] points out the irrelevance between IoU and classification predictions. Based on this, we adopt IoU between the final predicted mask and the ground truth as the objectness targets.

Sparse Activation takes the bilinear interpolation upsampling of the coarse feature map in the previous stage and the pixel-by-pixel summation of the feature map in the current stage as input. The result is then multiplied by the decision mask from the previous stage to obtain a sparse matrix, accelerating inference. During training, no sparsification is performed, and prediction is done with dense convolutions, similar to Dense Activation. During inference, we replace 3×3 dense convolutions with 3×3 sparse convolutions, utilizing the weights of the dense convolution to construct the sparse convolution [38] kernel. Finally, a residual structure is added to enhance contextual preservation.

In addition to the implicitly supervised activation described above, the injection of global features into the deeper features at AFD similarly complements the small target information in the deeper features of the network, which becomes the basis for the activation of the features in ACS. Enable the decision mask predicted by deep features to have the capability to guide the sparsification of shallow features.

C. LOSS FUNCTION

The training loss for ADSNet is the sum of the classification loss L_{cls}, IoU-aware prediction loss L_{IoU}, and the final mask segmentation loss L_{seg}:

$$L = L_{\text{cls}} + L_{\text{IoU}} + L_{\text{seg}} \tag{7}$$

We supervised IoU-aware learning using binary cross entropy loss, and both classification and final mask segmentation were supervised using focal loss [16].

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

Resnet has been shown to extract features with finite network depth [39], but the large sensory fields that are too deep in it are not suitable for small targets. In order to make the predicted decision mask suitable, we chose ResNet-18 as the downsampling backbone network, with output levels $\boldsymbol{B}_2$- $\boldsymbol{B}_4$ and output strides [4], [8], [16]. The experiment used Stochastic Gradient Descent (SGD) optimizer for 50,000 training iterations. The learning rate reached a base learning rate of 0.02 after 4,000 warm-up iterations and decreased tenfold after 30,000 and 40,000 iterations, respectively. All experiments were conducted on a computer with an Nvidia A4000 GPU and Intel i7-8550U CPU.

B. DATASETS AND METRICS

We evaluated our algorithm on the NUDT-SIRST and IRSTD-1k datasets. The NUDT-SIRST dataset consists of 1327 images with a size of 256×256 pixels. Approximately 37% of the images have at least 2 targets, and 96% of the targets meet the definition of small targets by the International Society for Optical Engineering: targets should be smaller than 0.15% of the entire image area. The IRSTD-1k dataset comprises 1000 images with a size of 512×512 pixels. The two datasets contain complex scenes such as sea and ground, with multi pose targets such as points, drones, and airplanes, simulating most of the actual detection scenarios. The ratio of training images to test images is set at 1:1.

In this paper, we choose standard performance metrics for infrared small target detection: *normalized IoU* (*nIoU*), Probability of Detection (*Pd*), False-Alarm Rate (*Fa*), and Receiver Operating Characteristic (ROC) curve. The *nIoU* is the normalized intersection over union between the predicted mask and the ground truth. The ROC curve visually reflects the relationship between the True Positive Rate (*TPR*) and False Positive Rate (*FPR*) in target detection. The *nIoU*, *TPR*, and *FPR* are defined as:

$$nIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{TP}{T+FP} \tag{8}$$

$$TPR = \frac{TP}{TP+FN} \tag{9}$$

$$FPR = Fa = \frac{FP}{N} \tag{10}$$

where TP, FP, FN, and T denote the number of detected true positive, false positive, false negative, and true target, respectively. Note that these are pixel level evaluation metrics. For pd we use target level to denote:

$$Pd = \frac{T_{\text{pred}}}{N_{\text{all}}} \tag{11}$$

where T_{pred} denotes the number of correctly predicted targets. N_{all} denotes the number of all target.

C. COMPARISON WITH STATE OF THE ART

We compare our method with SOTA infrared small target detection methods, including traditional algorithms: Top-hat [40], LEF [6], AADCDD [41], TLLCM [7]. Deep learning-based algorithms: ALCNet [4], DNANet [10], RDIAN [42].

TABLE 2. Comparisons with SOTA methods on NUDT-SIRST and IRSTD-1k in *nIoU*(%), *Pd*(%), *Fa*(10^{-6}).

Method	NUDT-SIRST			IRSTD-1k		
	nIoU	*Pd*	*Fa*	*nIoU*	*Pd*	*Fa*
Top-hat	18.12	78.99	848.01	10.24	75.10	1431.72
LEF	12.61	72.01	2063.13	7.17	83.63	689.51
AADCDD	14.47	67.43	107.49	8.49	66.18	86.51
TLLCM	7.34	75.45	1966.72	6.10	67.44	22.35
ALCNet	60.35	94.92	27.62	58.94	91.94	33.17
DNANet	**86.17**	97.35	**8.31**	63.13	92.05	15.26
RDIAN	83.98	94.39	14.17	59.46	88.43	27.14
Ours	85.94	**98.16**	9.42	**68.79**	**93.18**	**13.39**

The hyperparameter settings for traditional methods are shown in Table 1. To ensure the fairness of the experiments, all algorithms will be run under the same environment.

As shown in Table 2, the best values for each metric are indicated in bold, and the second-best values are underscored. ADSNet exhibits significant advantages over other state-of-the-art (SOTA) methods on both NUDT-SIRST and IRSTD-1k datasets. Since these two datasets encompass various complex backgrounds and target types, traditional algorithms are constrained by the prior features designed for specific scenarios, and the selection of hyperparameters also fails to achieve good generalization. The deep learning methods perform much better than the traditional methods, and the effect of controlling *Fa* is remarkable. Our method shows substantial improvement compared to other deep learning methods. This is attributed to AFD mapping multiscale features and ACS activating multi-instance awareness. ADSNet maintains a high *Pd* while keeping low *Fa*, with an *nIoU* on IRSTD-1k higher than similar methods by 5.66.

The ROC curves are shown in Fig. 5. Traditional methods fail to guarantee the *TPR* at low *FPR*, making it ineffective in filtering out background information. Deep learning methods, driven by image information, can achieve very low *FPR*. Our method has achieved good performance, demonstrating the effectiveness of the proposed method in suppressing the *FPR* while maintaining a high *TPR*.

The inference time of the deep learning algorithm for a single image in the NUDT-SIRST dataset is shown in Table 3, with lower values indicating better performance. For images in the NUDT-SIRST dataset with a size of 256^2, the average speed of inference for a single image using the proposed method is 0.041s. Due to the cascaded sparsification of high-resolution features, the control of computational cost is remarkable.

We visualize results on selected images with complex backgrounds from the two datasets, as shown in Fig. 6. Our method can segment targets in most complex background scenes. Top-hat focuses on high-frequency areas in the image and cannot distinguish between background object edges and small targets. LEF, AADCDD, and TLLCM have varying degrees of capture effect on targets by enriching local contrast judgment criteria, but also generate false positives due to prior

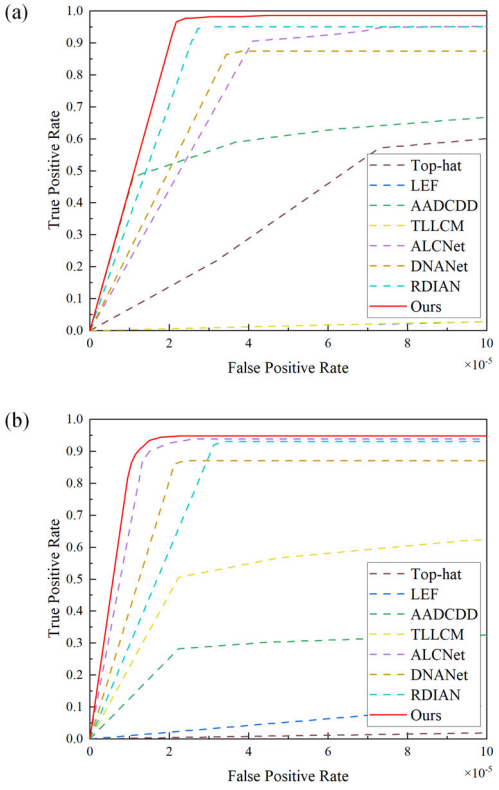

FIGURE 5. ROC curves of ADSNet and the SOTA methods, (a) ROC on NUDT-SIRST, (b) ROC on IRSTD-1k.

TABLE 3. Average inference time of single image for deep learning methods.

	ALCNet	DNANet	RDIAN	Ours
Time/image(s)	0.116	0.175	0.039	0.041

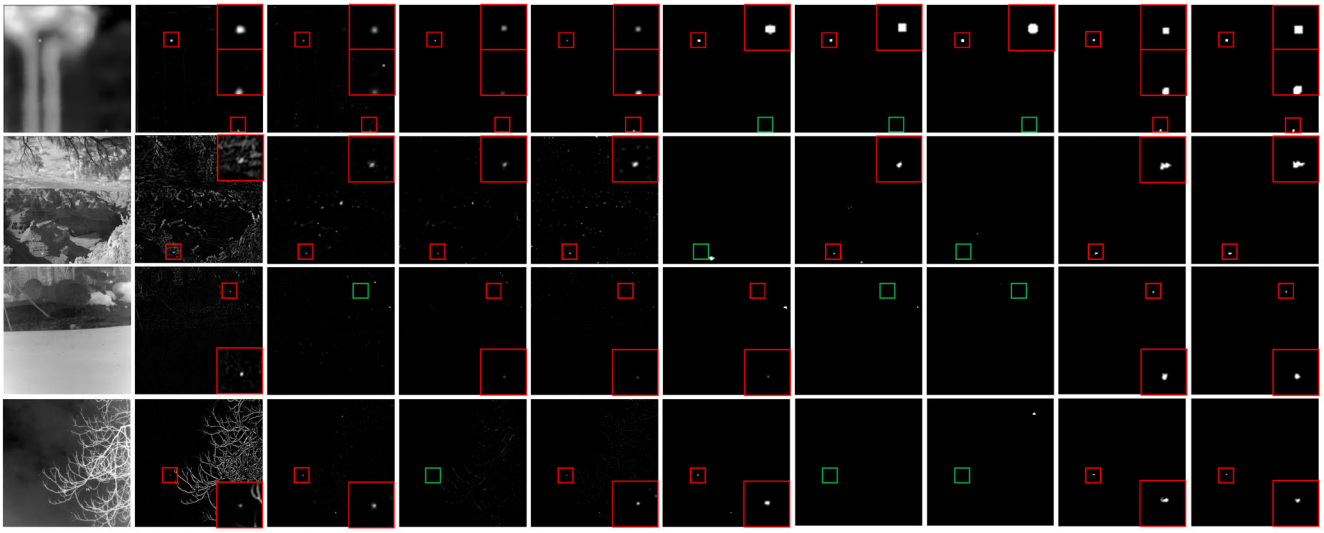

FIGURE 6. Visualization results for each method. Red boxes indicate that the target is detected and zoomed in, while green boxes indicate that the target is not detected.

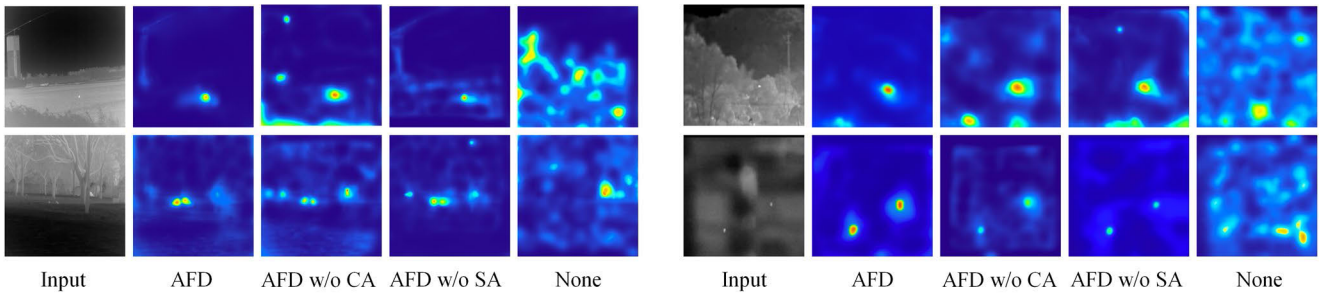

FIGURE 7. Visualization results of feature map fusion under different configurations of AFD.

models set in advance. In contrast, deep learning methods generate prediction probabilities at relevant positions and can effectively filter out background clutter. Our method aggregates multi-scale features, reduces feature vanishing, and utilizes multi semantic activation instance perception to enhance detection robustness, which is key to ensuring effective detection.

D. MODEL ANALYSES

1) IMPACT OF AFD MODULE

We visualize the $\boldsymbol{B}_4$ level, as shown in Fig. 7. With the assistance of both SA and CA, information can be provided to the deep-layer features of the network, ensuring the preservation of features for small infrared targets. As shown in Table 4, compared to the model without AFD, nIoU increased by 5.02%, PD increased by 5.35%, and Fa decreased by 4.01×10^{-6}. We further conducted ablation studies by individually ablating the two attention mechanisms, CA and SA, which brought different improvements to the overall detection performance. CA enhances the channel-wise information representation for small targets, while SA strengthens the position information of small targets, avoiding the introduction of background clutter similar to small targets, resulting in better detection results.

TABLE 4. Ablation study of the different configurations of AFD in *nIoU* (%), *Pd* (%), *Fa* (10−6).

Module	*nIoU*	*Pd*	*Fa*
-	80.92	92.81	11.43
AFD w/o CA	85.02	97.89	8.32
AFD w/o SA	84.66	97.65	8.57
AFD	85.94	98.16	7.42

We assigned global features to different down-sampling stages of the backbone network. As shown in Table 5, assigning $\boldsymbol{F}_{\text{global}}$ more to $\boldsymbol{B}_3$ resulted in an increase of 1.42% in *nIoU* and 0.96% in *Pd*, while *Fa* decreased by 0.89×10^{-6}. However, assigning $\boldsymbol{F}_{\text{global}}$ to $\boldsymbol{B}_{2-4}$ has limited improvement on the overall detection performance of the network but significantly impacts the inference speed. This is because most of the small target features are already present in the shallow features, and feature supplementation by $\boldsymbol{F}_{\text{global}}$ does

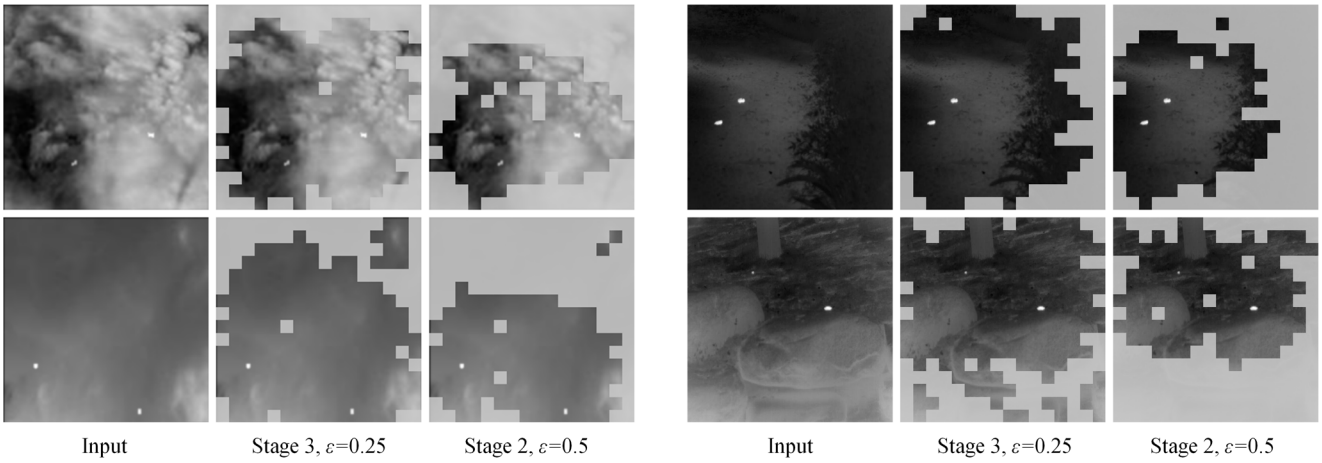

FIGURE 8. Visualization of the cascaded sparse patches on feature maps.

TABLE 5. Ablation study of distributing global feature to different stages in *nIoU* (%), *Pd* (%), *Fa* (10−6), Time.

Distribute to	*nIoU*	*Pd*	*Fa*	Time/s
Stage 4	84.52	97.20	8.31	0.038
Stage 3, 4	85.94	98.16	7.42	0.041
Stage 2, 3, 4	85.97	98.16	7.40	0.071

TABLE 6. Ablation on the IoU-aware in *nIoU* (%), *Pd* (%), *Fa* (10)-(6), Time.

IoU-aware	*nIoU*	*Pd*	*Fa*	Time/s
-	83.75	95.23	8.06	0.039
√	85.94	98.16	7.42	0.041

not result in a significant improvement, but rather slows down the inference speed due to the high resolution of the shallow features.

2) IMPACT OF ACS MODULE

As shown in Table 6, we investigated the improvement in overall detection performance introduced by IoU-aware supervision. Adding IoU-aware supervision increased *nIoU* by 2.19% without adding additional computational burden to the entire model.

We compared the detection performance of the network under different numbers of patches (i.e., resolution of decision mask). Before predicting the mask, the feature hierarchy was reshaped to the corresponding mask size. As shown in Table 7, the detection results gradually improved with an increase in the number of patches. Especially when the number of patches reached 16 × 16, the network obtains the highest accuracy.

TABLE 7. Ablation on the number of patches (resolution of decision mask) in *nIoU* (%).

Number of patches	*nIoU*
32×32	85.86
16×16	85.94
8×8	85.01
4×4	83.45

TABLE 8. Ablation on different sparse strategies in *nIoU*(%).

Sparse ratio ε		*nIoU*	Time/s
Stage3	Stage2		
0.2	0.4	85.97	0.056
0.25	0.5	85.94	0.041
0.3	0.6	84.73	0.032
-	0.25	86.05	0.071
-	-	86.12	0.096

As shown in Fig. 8, we conducted decision mask visualization. Patches covered by the mask are discarded. It can be observed that the decision mask gradually discards areas unrelated to the target, helping the network focus on identifying potential target locations while saving computational resources.

As shown in Table 8, we conducted an ablation study on mask sparsity ratio and sparsity position. We initiated sparsity from layer $\boldsymbol{F}_3$. Compared to the network without sparsity, *nIoU* decreased by 0.18, but the inference FPS increased by 13.9. Sparse activation on $\boldsymbol{F}_2$ alone similarly did not significantly improve the detection results. However, a larger sparsity ratio would compromise the overall network detection performance.

V. CONCLUSION

To accelerate the inference speed of infrared small target detection, we propose a cascaded sparse feature map scheme. However, information about small targets often diminishes as features are propagated layer by layer. To address this, we introduce the AFD module to complement deep network features. Simultaneously, we enhance the robustness of sparse feature map generation by using instance-aware activation. In our future work, we plan to apply ADSNet to conventional small target detection and explore its potential for accelerating the inference of sparse targets in 3D space.

REFERENCES

[1] T. Ma, Z. Yang, J. Wang, S. Sun, X. Ren, and U. Ahmad, "Infrared small target detection network with generate label and feature mapping," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3140432.

[2] 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, doi: 10.1109/TGRS.2020.3022069.

[3] 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.

[4] 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, doi: 10.1109/TGRS.2020.3044958.

[5] 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.

[6] 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, doi: 10.1109/LGRS.2019.2914432.

[7] 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, doi: 10.1109/LGRS.2019.2954578.

[8] Z. Zuo, X. Tong, J. Wei, S. Su, P. Wu, R. Guo, and B. Sun, "AFFPN: Attention fusion feature pyramid network for small infrared target detection," *Remote Sens.*, vol. 14, no. 14, p. 3412, Jul. 2022, doi: 10.3390/rs14143412.

[9] Y. Xu, Q. Chen, S. Kong, L. Xing, Q. Wang, X. Cong, and Y. Zhou, "Real-time object detection method of melon leaf diseases under complex background in greenhouse," *J. Real-Time Image Process.*, vol. 19, no. 5, pp. 985–995, Oct. 2022, doi: 10.1007/s11554-022-01239-7.

[10] B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, and Y. Guo, "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.

[11] 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. 3163410, doi: 10.1109/TGRS.2022.3163410.

[12] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 867–876, doi: 10.1109/CVPR52688.2022.00095.

[13] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 6890–6899, doi: 10.1109/ICCV48922.2021.00683.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[15] T. Cheng, X. Wang, L. Huang, and W. Liu, "Boundary-preserving mask R-CNN," in *Computer Vision—ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 660–676.

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.

[17] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Computer Vision—ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frah, Eds. Cham, Switzerland: Springer, 2020, pp. 282–298.

[18] T. Zhang, S. Wei, and S. Ji, "E2EC: An end-to-end contour-based method for high-quality high-speed instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4433–4442, doi: 10.1109/CVPR52688.2022.00440.

[19] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "PolarMask: Single shot instance segmentation with polar representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12190–12199, doi: 10.1109/CVPR42600.2020.01221.

[20] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask transfiner for high-quality instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4402–4411, doi: 10.1109/CVPR52688.2022.00437.

[21] Y. Pang, T. Wang, R. M. Anwer, F. S. Khan, and L. Shao, "Efficient featurized image pyramid network for single shot detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7328–7336, doi: 10.1109/CVPR.2019.00751.

[22] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," 2019, *arXiv:1908.05612*.

[23] Y. Liu, S. Cao, P. Lasang, and S. Shen, "Modular lightweight network for road object detection using a feature fusion approach," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 51, no. 8, pp. 4716–4728, Aug. 2021, doi: 10.1109/TSMC.2019.2945053.

[24] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212, doi: 10.1109/CVPR.2018.00442.

[25] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA-NET: Vertical spatial sequence attention network for traffic sign detection," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3423–3434, Jul. 2019, doi: 10.1109/TIP.2019.2896952.

[26] Y. Pang, J. Cao, J. Wang, and J. Han, "JCS-net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 12, pp. 3322–3331, Dec. 2019, doi: 10.1109/TIFS.2019.2916592.

[27] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13658–13667, doi: 10.1109/CVPR52688.2022.01330.

[28] X. Xie, G. Cheng, Q. Li, S. Miao, K. Li, and J. Han, "Fewer is more: Efficient object detection in large aerial images," *Sci. China Inf. Sci.*, vol. 67, no. 1, pp. 1–9, Jan. 2024, doi: 10.1007/s11432-022-3718-5.

[29] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787, doi: 10.1109/CVPR42600.2020.01079.

[30] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13034–13043, doi: 10.1109/CVPR46437.2021.01284.

[31] X. He, N. Tong, and X. Hu, "High-resolution I-imaging via MMV-based block-sparse signal recovery," *IET Radar, Sonar Navigat.*, vol. 13, no. 2, pp. 208–212, Feb. 2019, doi: 10.1049/iet-rsn.2018.5181.

[32] Z. Liu, L. Yu, and H. Sun, "Image denoising via nonlocal low rank approximation with local structure preserving," *IEEE Access*, vol. 7, pp. 7117–7132, 2019, doi: 10.1109/ACCESS.2018.2890417.

[33] Y. Dong, H. Che, M.-F. Leung, C. Liu, and Z. Yan, "Centric graph regularized log-norm sparse non-negative matrix factorization for multi-view clustering," *Signal Process.*, vol. 217, Apr. 2024, Art. no. 109341, doi: 10.1016/j.sigpro.2023.109341.

[34] Z. Chen, T. Wang, X. Wu, X.-S. Hua, H. Zhang, and Q. Sun, "Class re-activation maps for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 959–968, doi: 10.1109/CVPR52688.2022.00104.

[35] T. Cheng, X. Wang, S. Chen, W. Zhang, Q. Zhang, C. Huang, Z. Zhang, and W. Liu, "Sparse instance activation for real-time instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 4423–4432, doi: 10.1109/CVPR52688.2022.00439.

[36] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. Int. Conf. Learn. Represent.*, Nov. 2016, p. 5. Accessed: Nov. 27, 2023.
[37] S. Wu, X. Li, and X. Wang, "IoU-aware single-stage object detector for accurate localization," *Image Vis. Comput.*, vol. 97, May 2020, Art. no. 103911, doi: 10.1016/j.imavis.2020.103911.
[38] B. Graham and L. van der Maaten, "Submanifold sparse convolutional networks," 2017, *arXiv:1706.01307*.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 630–645.
[40] J. Rivest, "Detection of dim targets in digital infrared imagery by morphological image processing," *Opt. Eng.*, vol. 35, no. 7, p. 1886, Jul. 1996, doi: 10.1117/1.600620.
[41] 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, doi: 10.1016/j.infrared.2019.06.003.
[42] 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. 3235150, doi: 10.1109/TGRS.2023.3235150.

YANSONG SONG received the Ph.D. degree from Changchun University of Science and Technology, Changchun, China, in 2013. He is currently a Professor with the Institute of Space Ophotoelectronics Technology, Changchun University of Science and Technology. His current research interests include space laser communication technology and intelligent image processing.

BOXIAO WANG was born in Heilongjiang, China, in 1998. He received the B.S. degree in measurement and control technology and instrumentation from Changchun University of Science and Technology, Changchun, China, in 2021, where he is currently pursuing the master's degree in optical engineering. His main research interests include target recognition and localization image processing techniques.

KEYAN DONG is currently a Professor with the School of Optoelectronic Engineering, Changchun University of Science and Technology, China. His current research interests include space laser communication technology and computational optics.