Patch-Based Semantically Enhanced Network for IR Dim and Small Targets Background Suppression

Yunfei Tong, Yue Leng, Hai Yang, Zhe Wang, *Associate Member, IEEE*, Saisai Niu, and Huabao Long

***Abstract*—The task of background suppression in infrared small-target scenarios aims to eliminate irregular noisy backgrounds while preserving targets with high-frequency features. In infrared small-target scenes at long distances, the backgrounds become complex and the target features are degraded, highlighting a significant disparity between the detailed and realistic background and the limited features of the targets. To address these challenges, we propose a patch-based semantically enhanced generative adversarial network (GAN) named PSEnet for background suppression in infrared small-target scenarios. First, we introduce a patch-scale GAN that allows the model to concentrate on local background suppression. This shift from a global to local perspective simplifies the complexity of background suppression. Second, we employ the PSE module consisting multiscale dilated convolution and adaptive weight fusion to extract local semantic information. Third, by segmenting the infrared image into smaller patches and resampling them, we create a more balanced dataset for adversarial training. Experimental results demonstrate that the proposed algorithm significantly improves the signal-to-noise ratio of dim and small targets, reduces the missing detection rate, and achieves a precision of almost 91%. In conclusion, this approach effectively uses GANs for background suppression in complex environments.**

***Index Terms*—Background suppression, data imbalance, generative adversarial networks (GANs), multiscale feature fusion, low Signal-to-Noise Ratio (SNR) infrared (IR) scenes.**

I. Introduction

INFRARED (IR) imagery relies on variations in object radiation to capture detailed images, enabling it to penetrate through smoke, fog, dust, and snow, and identify camouflage, making it suitable for target detection in specialized environments. The objective of IR image background suppression is to eliminate extraneous sources of radiation in the background while preserving the high-frequency radiation of the target. This technique finds wide-ranging applications in military operations, biological observation, and other fields. However, in scenarios where early target detection is crucial, the increased distance between the scene and the camera results in more complex backgrounds and compromised target features. Specifically, the targets are often small in size and exhibit a low Signal-to-Noise Ratio (SNR) [1]. According to the Society of Photo-Optical Instrumentation Engineers (SPIE) [2], a target occupying less than 0.12% of the entire IR image and with an SNR below 5 dB is considered a dim and small target. Hence, detecting such targets with limited size and SNR presents significant challenges, typically encounters two ahead. First, the IR targets are generally small in size and with limited features. Second, the dim targets are often obscured by background clutter and significant noise, as shown in Fig. 1(b). Therefore, background suppression plays a vital role in enhancing dim and small targets in IR images, thus being essential for accurate target detection.

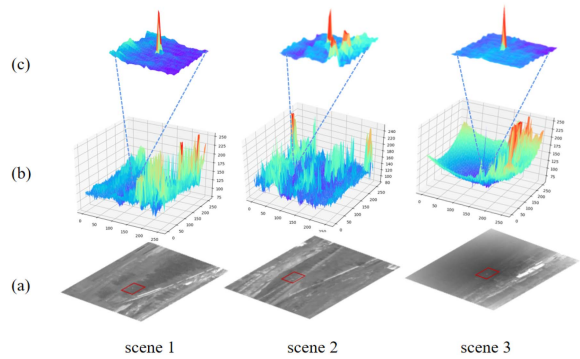

Fig. 1. Visualization of the dim and small IR images in three different scenes. (a) Original dim and small IR images. (b) 3-D gray-scale distribution map of the whole IR images. (c) 3-D gray-scale distribution map of the target area.

In the realm of IR dim and small-target background suppression, existing techniques can be classified into two distinct groups: 1) patch-based methods and 2) CNN-based methods [1]. These two categories encompass diverse approaches in addressing the aforementioned problem. Among patch-based methods, LCM [3], [4], [5] is the predominant approach, leveraging the human visual system (HVS) to examine local contrast information and mitigate clutter by manually defining target features and filtering conditions. But for complicated scenes, single-frame patch-based methods tend to fuse the targets with noise because of the low SNR. Other patch-based methods [6], [7], [8] utilize

Manuscript received 28 February 2024; revised 9 April 2024; accepted 24 April 2024. Date of publication 30 April 2024; date of current version 9 May 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3203500, in part by the Natural Science Foundation of China under Grant 62076094, in part by the Chinese Defense Program of Science and Technology under Grant 2021-JCJQ-JJ-0041, in part by the Shanghai Science and Technology Program "Federated based cross-domain and cross-task incremental learning" under Grant 21511100800, and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding authors: Hai Yang; Zhe Wang.)*

Yunfei Tong, Yue Leng, Hai Yang, and Zhe Wang are with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China, and also with the Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China (e-mail: yanghai@ecust.edu.cn; wangzhe@ecust.edu.cn).

Saisai Niu and Huabao Long are with the Shanghai Aerospace Control Technology Institute, Shanghai 201109, China, and also with the Research and Development Center of Infrared Detection Technology, China Aerospace Science and Technology Corporation, Shanghai 201109, China.

Digital Object Identifier 10.1109/JSTARS.2024.3394953

both spatial and temporal information to highlight dim targets and suppress background noise. Apart from patch-based methods, several CNN-based techniques [9], [10], [11], [12], [13] have been utilized for background suppression. For example, ACM [10], DNANet [11], and AGPCNet [12] employ the U-Net framework to generate background suppression images while MDFAcGAN [13] leverages an adversarial generative network to learn the distinctions between the targets and background. With the rise of deep learning, data-driven training methods have gained popularity. Prior research findings demonstrate that CNN-based methods exhibit superior performance in terms of both positive detection and false alarm rates. However, the performance of data-driven approaches largely depends on the quality of the training data. IR small-target imagery typically features complex backgrounds with limited target characteristics, creating a significant imbalance [14]. To address this, some methods utilize data augmentation or introduce weighted loss functions. For balanced sampling, oversampling [15] is a straightforward yet efficacious approach that is well-established in handling unbalanced classification problems. Building on this, DS-GAN [16] offers a comprehensive data augmentation pipeline for enhancing small-target detection. It generates a diverse set of small-target samples through generative adversarial networks (GANs), which require high-resolution target images. However, such high-resolution data are scarce for IR small-target detection scenarios. On another hand, weighted classification losses, such as focal loss [17], may help improve performance, but their effectiveness can be compromised by parameter sensitivity and may not be fully effective in extremely unbalanced situations with sparse small targets. Therefore, accurately distinguishing targets with fewer pixels and limited features from global images remains a challenging task.

Although the aforementioned methods have brought some improvements to the image enhancement of dim and small IR targets, they also have two problems. First, the unbalanced distribution of positive and negative samples arises due to the small size of the target area. Second, the dim and small targets often lack distinctive texture features and exhibit similar characteristics to the background noise.

In order to address the issue of data imbalance, the select patch (SP) module is designed for data preprocessing. This module enables random sampling from the IR images to achieve a balanced representation of target area patches and background area patches. Without the SP strategy, the image pixels are predominantly of the background class, resulting in a scarcity of positive samples. Consequently, the data imbalance problem is effectively resolved. In addition, the SP model augments the number of samples available for the network to learn from, thereby serving the purpose of expanding the dataset.

To solve the problem of noise interference through the process of background suppression, we designed a Patch-based Semantically Enhanced Network (PSEnet). As depicted in Fig. 1, target patches exhibit consistent distributions across different scenes while IR images from distinct scenes tend to differ primarily due to the dominant background. The target is typically closely associated with the local background and shows minimal correlation with the distant background within the image. Rather than focusing on determining the presence of a target within an entire image, it is more advantageous to learn the distributional disparities between the target patches and the background patches at a finer granularity. Based on that, we use the spatial and semantic information of patches to obtain the local correlation of features. Due to patch-based learning, the network can better learn the local information of the whole image. Specifically, the overall network adopts a GAN structure. The generator is designed to better obtain the feature representation of dim and small targets, possessing the following three characteristics: using a shallow CNN network, fusing shallow spatial feature maps with deep semantic feature maps and increasing the target receptive field by dilated convolution. Moreover, as the network focuses on the subtasks of segmenting dim and small objects from local IR images, the overall difficulty of background suppression is decreased. At last, the local-to-global (LTG) module is designed to get the overall suppression result from the patch to the whole image.

In general, the entire dim and small targets enhancement procedure can be described in three steps. First, put the original images and labels into the SP module to select the local patch. Then, suppress background by the PSEnet. Finally, use the LTG module to transfer the local background suppression result to the overall. In summary, we propose a patch-based adversarial learning paradigm to solve the difficulties of data imbalance and noise interference caused by dim and small targets. The contributions of this article can be summarized as follows.

1) The SP module is proposed to address the challenge of imbalanced data in small-object image enhancement. The SP module plays a pivotal role in this task, serving two main purposes. First, it alleviates the issue of data imbalance by ensuring an equal number of positive and negative samples. Second, it enhances the model's focus on local spatial information, leading to improved performance in enhancing small objects.
2) To achieve superior enhancement of IR small-target images, we propose the GAN-based PSEnet framework. This framework is designed to suppress the background, enhance the SNR, and amplify the intensity of the target. Specifically, we introduce the PSE module, which enables progressive multilayer feature fusion and texture enhancement. Furthermore, the integration of the LTG module facilitates the efficient propagation of local results throughout the entire image context.
3) PSEnet achieves the best segmentation results on nine sequences. The results prove that the background suppression method can not only improve the SNR of the dim and small targets but also decrease the missing detection rate in the complex scenes.

II. Related Work

In this section, we briefly review the major works in patch-based learning and CNN-based learning for IR dim and small targets enhancement and background suppression.

A. Patch-Based Learning for IR Dim and Small Targets Enhancement

Based on the HVS, Chen et al. [3] proposed the LCM, which calculates the local contrast of the image and uses the contrast feature to enhance the target area and suppress the background. Following LCM, WLDM [18] proposes a weighted-local-difference-measure-based scheme to simultaneously enhance targets and suppress background clutters and noise. AMWLCM [4] simultaneously exploits the local contrast of the target, the consistency of the image background, and the imaging characteristics of the background edges. HBMLCM [5] transformed the background suppression task into an optimization problem. Different from local-contrast-based methods, some patch-tensor-based methods [19], [20], [21], [22], [23], [24] enhance and detect an IR dim and small object by different strategies. For example, Top-Hat [19] and Max-Mean [20] reduce background clutter, and Qin et al. [21] applies the facet kernel. MPCM [23] utilizes the mean difference of different directions in multiscale patches to improve the stability and noise immunity of the algorithm. LogTFNN [22] designs a new IR patch-tensor model to have a better representation of background rank and robustness against noise interference. However, the aforementioned methods perform poorly when the targets are submerged in intricate clutter. To address this, the local energy factor (LEF) [25] conceives a local dissimilarity descriptor to enhance targets. Other methods [26], [27], [28], [29], [30], [31] enhance the target and suppress the background by optimizing the sum minimization of patch singular values.

B. CNN-Based Learning for IR Dim and Small Targets Enhancement

Traditional patch-based methods are always based on strong prior assumptions about the IR small targets and only use gray-scale values as features regarding the difference of semantic context between the targets and the background. Generative networks possess the ability to transform images or random noise into various image styles by employing an encoding and decoding structure. When combined with pixel-level loss constraints, segmentation networks often leverage generative architectures to generate binarizationlike probability maps. DNANet [11] achieved progressive interactions between high-level and low-level features by designing a dense nested interaction module. Semantic supplementary network [32] used the dependency relationship between labels to improve recognition accuracy. IAANet [1] applied a region proposal network to obtain rough object regions and filter out the background. AGPCNet [12] proposed an attention-guided context block, a context pyramid module, and an asymmetric fusion module to enhance the utilization of features. GAN, as a subtype of generative networks, incorporates a discriminator that employs an adversarial training model to produce images that are more realistic or aligned with the target style. The advantage of GANs lies in the emphasis on generating realistic edges and capturing fine details in the synthesized images. Moreover, it exhibits improved generalization by learning the distribution of IR images more effectively. To enhance target images, IE-CGAN [33] generated images with enhanced contrast and details using a fully convolutional network. MoCoPnet [34] integrated the domain knowledge of IR small objects into the deep network to alleviate the inherent feature scarcity of IR small objects. EESRGAN [35] applied superresolution networks to small-object edge enhancement, improved the quality of small objects, and used different detector networks for end-to-end operations. MDvsFAcGAN [13] uses two generators to get a balanced high precision rate and missing alarm rate. However, the adversarial structure's drawback could be the potential training challenges in achieving a balance between the generator and the discriminator.

III. Method

A. Method Overview

IR small-target scene images exhibit distinct distributions, as illustrated in Fig. 1(b). However, across different scenes, the local patches surrounding the target region may display similar distributions, as depicted in Fig. 1(c). Notably, compared to the entire IR image with or without small targets, the differences between the image patches with and without targets are more significant. Considering that GANs are generative models designed to learn data distributions, it is advantageous to focus on learning patch-scale data distributions. Thus, in this article, we propose novel background suppression methods that emphasize acquiring localized patch information instead of analyzing the entire IR image. By adopting this approach, we aim to achieve a high detection rate and a low false alarm rate.

The overflow of our method for IR dim and small-target enhancement is shown in Fig. 2. It can be summarized as three steps. First, the whole IR images and labels are divided into the target-area part containing the target and the background part without the target area through the SP module to obtain balanced positive and negative samples of the image patch and label patch. Second, the image patch and label patch are put in the PSEnet to get the gen-patch (the local background suppression results). The PSEnet is committed to learning the mapping relationship between the image patch and label patch. Third, the gen-patch is put into the LTG module to get the segmentation results of the entire IR dim and small-target images.

B. SP Module

One difficulty of the IR image in dim and small-target detection is that target pixels only occupy an extremely small proportion of the image, which means that most parts of the image are redundant. As a result, the network does not have enough positive samples during training for the network to learn how to distinguish these with noise points. So, the article proposes the SP model before segmenting. The main procedure of the SP module is shown in Fig. 3.

The SP module takes as input an original IR image raw_{IR}, its corresponding binary annotation image with the same size $\text{raw}_{\text{binary}}$, and the target center (t_x, t_y). The size of the sampling block sz and the number of positive and negative sample pairs st per image need to be specified. To obtain matching IR image patches and binary label patches for each sample, the same

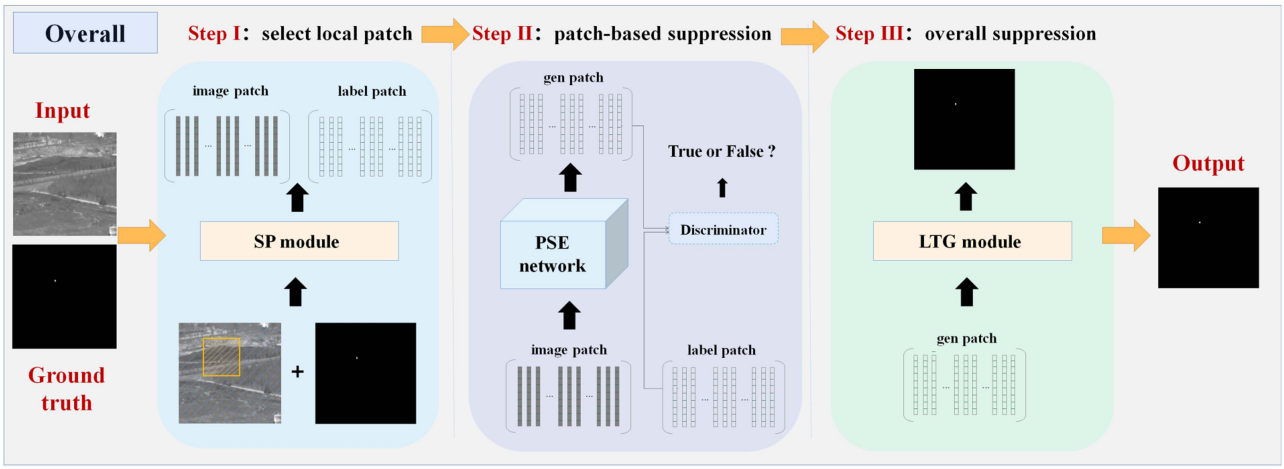

Fig. 2. Overall flow of our method. In the initial stage, image patches are randomly cropped from the training set, maintaining a 1:1 ratio between background and target samples. Subsequently, adhering to the adversarial generation paradigm, the generator within PSEnet produces suppression results, which are subsequently assessed by the discriminator to ensure alignment with the labeled image. During the inference phase, the original image is partitioned into patches, and the corresponding suppression results are acquired. Finally, these results are synthesized into a unified image using the LTG module.

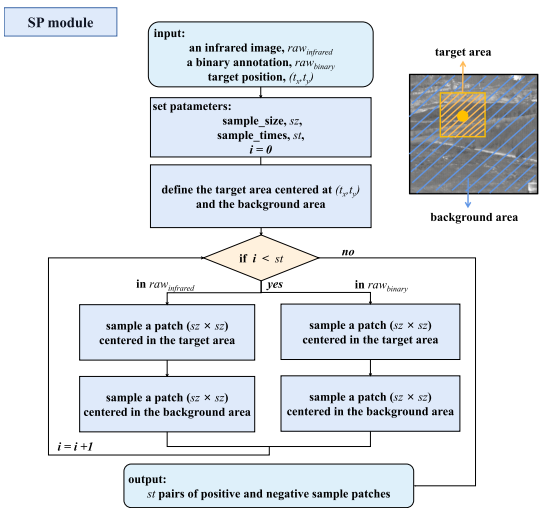

Fig. 3. Overflow of our proposed SP module.

operations are performed on both inputs. First, the target region is defined as an image patch with a center at (t_x, t_y) and a side length of sz. The remaining area in the image is referred to as the background region. Second, the center positions (b_x, b_y) for the background patch are randomly determined within the image. Third, image patches of size $sz \times sz$ are cropped around the centers (t_x, t_y) and (b_x, b_y) from both the input image and the annotation image. This yields one positive and one negative sample patch. This process is repeated n times on the same image, resulting in st positive and st negative samples. By performing this operation on every image in the dataset, a balanced training dataset is achieved, with st positive and st negative samples per image. The sample size and sample times were determined through experimental analysis. In Section IV, this article investigates the impact of patch size (sz) using values of 16, 32, and 64. The experimental results reveal that setting sz to 32 yields the best performance. In addition, the sample times were determined based on precision and resource consumption considerations. To achieve superior outcomes with reduced resource usage, we set the sample times (st) to 50. Subsequently, the SP module randomly selects 100 image patches and 100 label patches from both the background and target areas. Each set comprises 50 positive samples and 50 negative samples. This careful selection process ensures a balanced representation of samples for comprehensive analysis and evaluation.

C. Patch-Based Small Object Segmentation Network

The second step of the proposed method involves patch-based semantic enhancement (PSE) of IR images using PSEnet after obtaining a balanced set of target patches and background patches through the SP module. The PSEnet network structure, depicted in Fig. 4, is based on generative adversarial architecture. However, the low input image resolution and sparse feature representation of small targets make it challenging to use traditional object segmentation methods directly, as they tend to obscure the features of small targets and result in high rates of missed detections and false positives. To address this issue, the proposed method employs a relatively shallow feature extraction network structure designed to enable the model to focus on low-level spatial structural information, guiding downstream target detection tasks and improving small-target detection. PSEnet utilizes the PSE module and local residual structure to propagate information on small targets.

The purpose of using dilated convolutional layers in the PSE module is to expand the network's receptive field without increasing the number of parameters, thereby reducing the loss of internal structure and position of targets caused by subsequent pooling layers. Specifically, the PSE module employs dilated convolutions with different dilation rates to obtain contextual information with varying receptive fields, thereby enhancing the local semantic context of the image. The structure of the module, as shown in Fig. 5, first processes the input feature

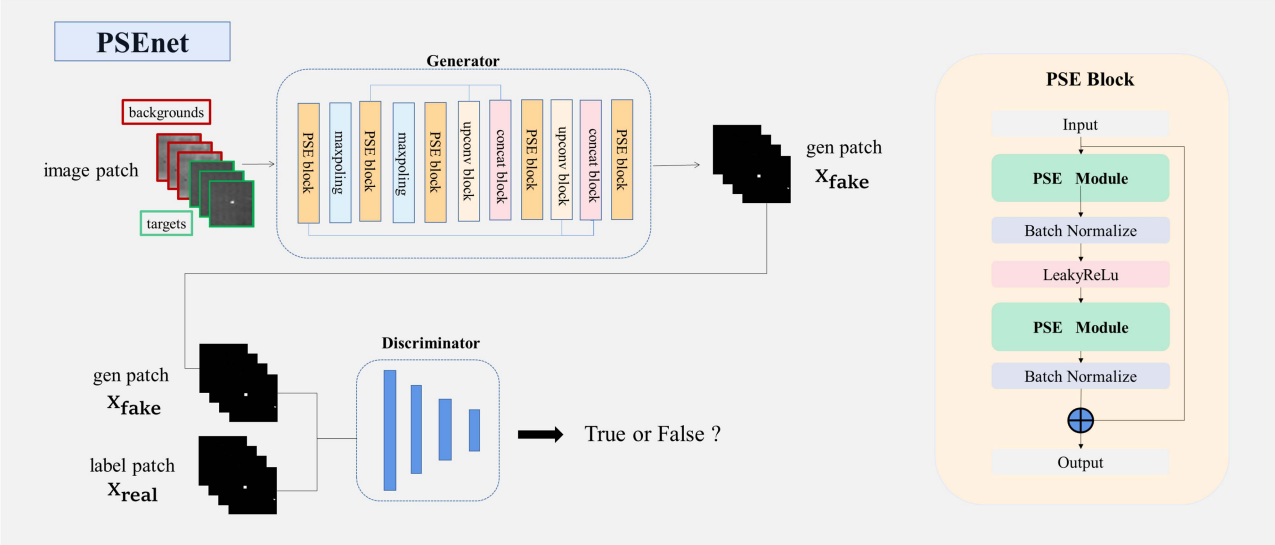

Fig. 4. Architecture of our proposed PSEnet network. PSEnet adheres to the adversarial generative network paradigm and comprises a generator and a discriminator. Specifically, our designed generator focuses on small targets and employs a shallow CNN network. It increases the receptive field through dilation convolution and combines the shallow spatial feature map with the deep semantic feature map.

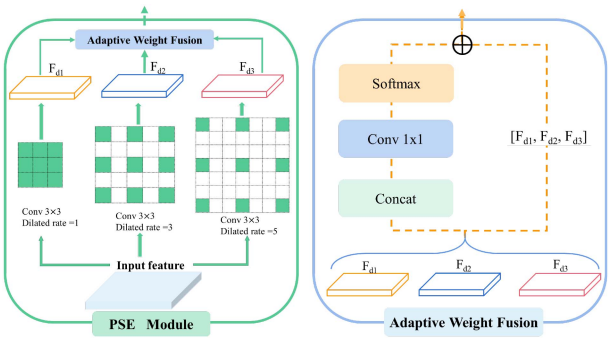

Fig. 5. Architecture of the PSE module in the PSEnet network.

maps using dilated convolutions with dilation rates of 1, 3, and 5, respectively, and then adapts to different target scales through adaptive feature fusion. The input feature maps $\mathbf{F}_{\mathbf{d1}}^{(i)}$, $\mathbf{F}_{\mathbf{d2}}^{(i)}$, and $\mathbf{F}_{\mathbf{d3}}^{(i)}$ with the same channels are concatenated before being compressed to 3 channels using a 1×1 convolution. The softmax operation is then applied to assign feature weights to the three channels. Overall, the computational process of the PSE module can be represented as follows:

$$\mathbf{F}_{\mathbf{d1}}^{(i)} = \text{dilated}_{1\times1}(\mathbf{F}^{(\mathbf{i})}) \tag{1}$$

$$\mathbf{F}_{\mathbf{d2}}^{(i)} = \text{dilated}_{3\times3}(\mathbf{F}^{(\mathbf{i})}) \tag{2}$$

$$\mathbf{F}_{\mathbf{d3}}^{(i)} = \text{dilated}_{5\times5}(\mathbf{F}^{(\mathbf{i})}) \tag{3}$$

$$\lambda_{\alpha^{(i)}}, \lambda_{\beta^{(i)}}, \lambda_{\gamma^{(i)}} = \text{Conv}_{1\times1}\left[\text{Concat}\left(\mathbf{F}_{\mathbf{d1}}^{(i)}, \mathbf{F}_{\mathbf{d2}}^{(i)}, \mathbf{F}_{\mathbf{d3}}^{(i)}\right)\right] \tag{4}$$

$$\alpha^{(i)}, \beta^{(i)}, \gamma^{(i)} = \frac{e^{\lambda_{\alpha^{(i)}}}, e^{\lambda_{\beta^{(i)}}}, e^{\lambda_{\gamma^{(i)}}}}{e^{\lambda_{\alpha^{(i)}}} + e^{\lambda_{\beta^{(i)}}} + e^{\lambda_{\gamma^{(i)}}}} \tag{5}$$

$$\mathbf{F}^{(\mathbf{i})*} = \alpha^{(i)} * \mathbf{F}_{\mathbf{d1}}^{(i)} + \beta^{(i)} * \mathbf{F}_{\mathbf{d2}}^{(i)} + \gamma^{(i)} * \mathbf{F}_{\mathbf{d3}}^{(i)}. \tag{6}$$

Here, $\text{dilated}_{1\times1}$, $\text{dilated}_{3\times3}$, and $\text{dilated}_{5\times5}$ represent dilated convolutional layers with dilation rates of 1, 3, and 5, respectively.

To extract more informative features during the upsampling process, PSEnet combines the feature of the second convolution block and the first upsampling convolution block through splicing. Furthermore, the unique diagnostic information from the first convolution block is fused into the second upsampling layer. The generated positive and negative sample images, along with positive and negative labels, are then input into the discriminator to enable it to learn to distinguish between real and fake images. The discriminator plays a crucial role in distinguishing between real segmentation labels and fake suppression results produced by the generator. For the discriminator, we employ four fully connected layers for fitting, and the final output is obtained through the sigmoid function.

To better optimize the block-based IR small-target image enhancement network PSEnet, the loss function consists of three components.

Adversarial loss: The discriminator is trained to become better at making this distinction, thus forcing the generator to improve its output to create more convincing results. This competitive dynamic between the discriminator and the generator drives the improvement of the quality of the generated images over time. The two loss functions are described as follows:

$$D\_Loss_a = \mathbb{E}_{\mathbf{X_{real}}}(-\log(D(\mathbf{X_{real}}))) \tag{7}$$

$$+ \mathbb{E}_{\mathbf{X_{fake}}}(-\log(1 - D(\mathbf{X_{fake}}))) \tag{8}$$

$$G\_Loss_a = -(1 \times \log(D(\mathbf{X_{fake}}))) + 0 \times \log(1 - D(\mathbf{X_{fake}})) = -\log(D(\mathbf{X_{fake}})) \tag{9}$$

where $\mathbb{E}$ refers to the expectation of a set, $\mathbf{X_{real}}$ denotes the true segmentation label, and $\mathbf{X_{fake}}$ denotes the generated result. $D(\cdot)$ refers to the discriminator.

Data loss: If only relying on the discriminator to constrain the generation of the generator is not enough, so we add a data loss to calculate the structural similarity between the generated map and the labeled map and hope that the image generated by the generator can be better than the labeled map. Close structural similarity, we use L1 Loss to calculate the structural similarity between the generated image and the labeled image. The specific formula is described as follows:

$$G\_Loss_{\text{data}} = |\mathbf{X_{fake}} - \mathbf{X_{real}}|. \tag{10}$$

Object loss: To accurately suppress the background and ensure semantic consistency between the generated background suppression result and the original input IR image, we introduce an object loss that enforces pixel-level constraints on the generated map. As our objective is to differentiate the target from the background using the target loss, we employ the binary cross-entropy loss, represented as (11). By incorporating the object loss, we encourage the network to accurately identify and preserve the target region while suppressing irrelevant background information.

$$\begin{aligned} G\_Loss_{\text{obj}} = & -\mathbf{X_{real}}\log(\mathbf{X_{fake}}) \\ & -(1-\mathbf{X_{real}})\log(1-\mathbf{X_{fake}}). \end{aligned} \tag{11}$$

Total loss: Finally, the previous losses are combined with different coefficients, where λ_a represents the coefficient of adversarial loss, λ_{data} represents the coefficient of data loss, λ_{obj} represents the coefficient of obj loss. Take the values 1, 100, and 1, respectively, to update the discriminator and generator

$$L_G = \lambda_a G\_Loss_a + \lambda_{\text{data}} G\_Loss_{\text{data}} + \lambda_{obj} G\_Loss_{\text{obj}} \tag{12}$$

$$L_D = \lambda_a D\_Loss_a. \tag{13}$$

D. LTG Module

The LTG module plays a crucial role in the proposed method as it facilitates the mapping from local image patches to the whole image, enabling seamless integration. During the training phase, positive and negative samples are randomly selected using the SP model. These samples are then processed by the PSEnet to generate enhanced local patches. Subsequently, these enhanced patches are fed into the LTG module, where they are seamlessly integrated to form a complete and enhanced image, as depicted in Fig. 6. In the LTG module, the input consists of probability maps generated by the pretrained PSEnet network. First, the mean (μ) and variance (σ) of the background and target patches within each local region are computed by (17) and (18), respectively,

$$\mu = \frac{\sum_m^{sz}\sum_n^{sz} x_{m,n}}{sz \times sz} \tag{14}$$

$$\sigma = \sqrt{\frac{1}{sz \times sz}\sum_m^{sz}\sum_n^{sz}(x_{m,n}-\mu)}. \tag{15}$$

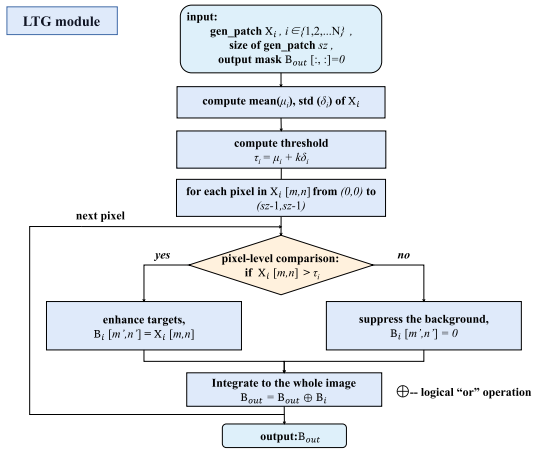

Fig. 6. Overflow of our proposed LTG module.

Second, the adaptive threshold (τ) is calculated based on (19)

$$\tau = \mu + k\sigma. \tag{16}$$

If the image confidence exceeds the threshold, the corresponding positions of the output are filled with 1, indicating the target output. On the other hand, if the patch output falls below the threshold, it is set to 0. This approach helps to improve the robustness of the target segmentation results and ensure that the output is accurate and reliable. Finally, the integrated image is output after applying Gaussian filtering, which further enhances the overall image quality.

IV. Experiments

A. Datasets

The dataset used in our experiment is from the work in [36], which covers the sky, the ground, and other scenes and contains a total of 22 segments of data, 30 tracks, 16 177 frames of images, and 16 944 targets. Each target size is different because of the distance. According to the SPIE [2], the two attributes of the targets, we selected the following nine sequences from the original dataset according to the difficulty of the target SNR and divided these nine sequences into simple scenes, medium scenes, and complex scenes.

We described the nine sequences in detail (see Table I). Among them, "data2," "data16," and "data19," are easy scenes, and we select "data19" as the testing sequence. As we can see in Fig. 7, "data2" has two targets flying across in the background. "data16" has a single target flying from near to far in the background. "data19" has a single target flying maneuverable in the background. Moreover, "data6," "data8," and "data18" are middle scenes, and we choose "data8" to be the testing sequence. Looking at Fig. 7, "data6" and "data8" have a single target flying from near to far in the ground background. "data19" has a single target flying in the background. Finally, "data11," "data12," and "data14" are difficult scenes, and "data11" is selected for testing. About Fig. 7, "data11" and "data14" have a single target flying from near to far in the ground background. "data12" has a single target flying from far to near in the background.

TABLE I
DESCRIPTION OF OUR DATASETS

Name	Size$_m$	Frames	Describe
data2	5	599	two targets, sky background, cross flight
data6	2	399	from near to far, single target, ground background
data8	2	399	from near to far, single target, ground background, for testing
data11	3	745	from near to far, single target, ground background, for testing
data12	2	1500	from far to near, single target, mid-target maneuver, ground background
data14	3	1426	from near to far, single target, ground background, interfered by ground vehicle target
data16	6	499	from near to far, single target, extended target, target maneuver, ground background
data18	3	500	from far to near, single target, ground background
data19	3	1599	single target, target maneuver, ground background, for testing

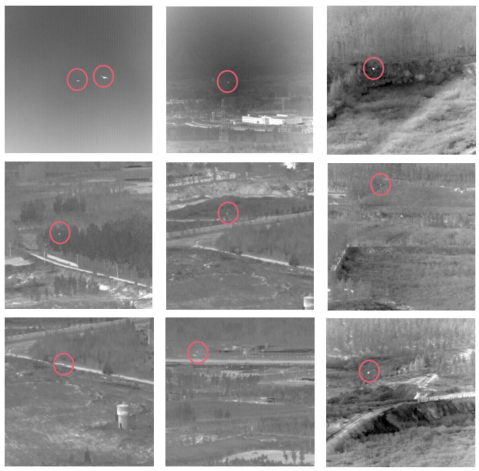

Fig. 7. Example images of three scenarios. The first row is a simple scene, the second row is a medium scene, and the third row is a complex scene. At the same time, the first two columns are used as the training set, and the last column is the testing set.

We randomly select 80% of the pictures (about 400 frames) from the training sequence to be the training data for the small object enhancement network. The rest of the pictures are used for testing. Finally, we used the pictures in the testing sequence to test the effect of the background suppression.

B. Experimental Settings

The experiment was conducted on a computer with a 2.50 GHz CPU, 8 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU. Our model is implemented in Python and Pytorch. We use the Adam algorithm for optimization. The key parameters are empirically determined to be $\alpha_1 = 0.5$ and $\alpha_2 = 0.999$, and they are applied uniformly to all experiments. The minibatch size is set to 128. The learning rate is set to be 2e-4 for the generators and the discriminator. The weights of the generator will be updated after the discriminator updates five times, and the whole training process terminates in 50 epochs.

To evaluate the background suppression and target enhancement effects of the IR images, the SNR is usually used as evaluation indicators. The higher the SNR of a small target, the easier it is to detect.

$$\mathrm{SNR} = (E_r - E_B)/\delta_B. \tag{17}$$

Among them, E_r is the mean value of the target area, E_B is the mean value of the background area, and δ_B is the standard deviation of the background area. Generally, the size of the background area is three times the size of the target area. $\mathrm{SNR_{in}}$ represents the SNR of the input images, and $\mathrm{SNR_{out}}$ represents the SNR of the output images. C_{in} and C_{out} are the standard deviations of the input image and enhanced image.

Besides, to illustrate that our method can effectively improve the local SNR of IR dim and small targets, we also add the image LSNR to verify the effectiveness of our experiments.

$$\mathrm{LSNR} = 10 \times \log_{10}\mathrm{SNR}. \tag{18}$$

For comparison, we use precision and recall as evaluation of object segmentation and compare each pixel of the enhanced binarized image result with the real value. If they are the same value, then regarded as a positive example, otherwise regarded as a negative example, and its calculation method is as follows:

$$\mathrm{precision} = \frac{\text{Number of positive samples}}{\text{Number of true targets}} \tag{19}$$

$$\mathrm{recall} = \frac{\text{Number of positive samples}}{\text{images}}. \tag{20}$$

In order to better evaluate the results of image enhancement, we use F1 to balance the relationship between accuracy and recall. The specific calculation formula is as follows:

$$\mathrm{F1} = \frac{2 \times \mathrm{recall} \times \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}}. \tag{21}$$

Mean average precision (mAP) is a popular evaluation metric for object detection tasks. It combines precision and recall to provide a single performance score. In the case of a single-class scenario, the formula for calculating mAP is as follows:

$$\mathrm{mAP} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{AP}_k. \tag{22}$$

Here, n represents the number of thresholds used, and AP_k denotes the average precision computed at the kth threshold.

The experiment mainly compares two kinds of methods, one is the patch-based IR small-target background segmentation method, and the other is the CNN-based IR small-target segmentation method. The article uses the pioneering LCM [3], Tophat [19], MPCM [23], AMWLCM [4], and LEF [25] to compare with the proposed methods. In CNN-based schemes, the article uses ALCnet [9], MDvsFAcGAN [13], AGPCnet [12], and IAAnet [1] as comparison methods. The parameter settings involved in these methods are given in Table II.

TABLE II
PARAMETERS OF METHODS FOR COMPARISON

Methods	Parameter Setting
Tophat	Structure shape: disk, Size: 5×5
LCM	Filter radius: 1, 2, 3, 4
MPCM	Mean filter size: 3×3, k=0.5
AMWLCM	Mean filter size: 5×5, k=0.8
LEF	Mean filter size: 5×5, $\alpha = 0.5$, $h = 0.2$

TABLE III
ABLATION STUDY ON PSENET AND LTG MODULE

method	data8	data11	data19
	F1(P,R)	F1(P,R)	F1(P,R)
PSEnet+LTG	0.52(0.77,0.30)	0.52(0.66,0.43)	0.43(0.51, 0.32)
PSEnet*+LTG	0.60(0.90,0.45)	0.61(0.72,0.57)	0.60(0.73, 0.52)
PSEnet*+LTG†	**0.94(0.95,0.92)**	**0.82(0.93,0.75)**	**0.77(0.78,0.75)**

"*" represents the network with dilated convolution. "†" represents using the adaptive threshold instead of the solid threshold. The experiments are conducted on data8,11,19 with F1. The best performance is in boldface.

C. Ablation Study

In the ablation study, we investigated several crucial aspects to comprehend the contributions of our model components. First, we examined whether the proposed PSEnet and LTG models can independently enhance background suppression from local and global perspectives. Second, to validate the effectiveness of the PSE block for small object detection, we compared it with existing feature extraction blocks and conducted experiments. Third, we explored the optimal values for the sample times and sample size of the SP module. Finally, we investigated the impact of integrating our method before detection on the accuracy rate of the detection results. By addressing these questions, we gained valuable insights into the effectiveness and potential of our model in improving background suppression and detection accuracy.

To ascertain the efficacy of the proposed PSEnet and LTG models in enhancing background suppression, we conducted experiments with different settings on PSEnet and LTG. To show the improvement by dilated convolution layers, we compare the PSEnet with the convolution layers and with the dilated convolution. The results in Table III show the dilated convolution can improve the accuracy of detection because it can increase the reception field of the target feature. To show the improvement by using of adaptive threshold in the LTG module, the article sets the parameters of solid and adaptive. The dilated convolution results are presented with "*," and the adaptive threshold results are presented with "†,"

To validate the effectiveness of the PSE block for small object detection, we compared it with existing feature extraction blocks and conducted experiments. According to Table IV, the utilization of the proposed PSE module demonstrates higher accuracy rates compared to other modules. This can be attributed to several factors, which are as follows:

1) the adoption of a shallow feature encoding structure to minimize the loss of target feature representation during feature pooling;
2) the enhancement of target context information through dilation convolution, enabling the model to access higher-level semantic information at an early stage;
3) the incorporation of multiscale feature fusion with adaptive weighting to accommodate weak targets of varying sizes.

TABLE IV
OBJECT DETECTION RESULTS UNDER DIFFERENT FEATURE ENHANCEMENT MODULES

method	data8	data11	data19
	F1(P,R)	F1(P,R)	F1(P,R)
Res block	0.89(0.92,0.84)	0.75(0.84,0.52)	0.63(0.71,0.54)
Attention block	0.93(0.94,0.92)	0.80(0.88,0.68)	0.75(0.77,0.72)
PSE block	**0.94(0.95,0.92)**	**0.82(0.93,0.75)**	**0.77(0.78,0.75)**

Best results are shown in bold.

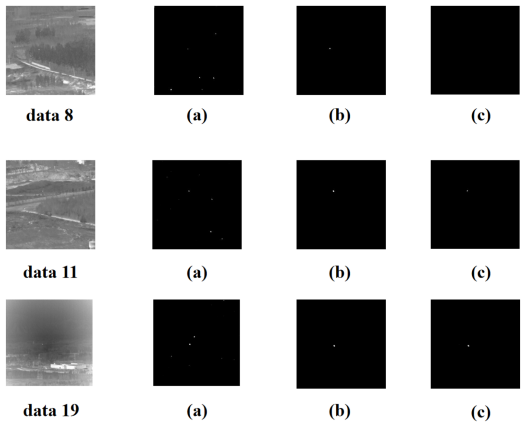

Fig. 8. In the experiment, in three scenarios, changing the size of the sampling frame (a) 16×16, (b) 32×32, and (c) 64×64, the results obtained, respectively.

Consequently, the PSE module proves to be better suited for the shallow structure background suppression network outlined in this section when compared to the residual structure and attention mechanism. Notably, when compared to the Res block and Attention block, the PSE block yields improvements of 0.05, 0.07, and 0.14 in the F1 metrics for data8, data11, and data19, respectively, thus validating the effectiveness of the proposed module.

To strike a balance between the model's processing efficiency and its generation effect, ablation experiments are conducted by altering the size of the sampling frame and the number of sampling iterations. When the sampling frame is larger, the small-sized targets occupy a relatively smaller proportion of pixels within it, causing the enhanced network to struggle in effectively segmenting the targets from the background. Conversely, if the sampling frame is too small, although the proportion of target pixels within the sampled image increases, the reduced coverage of background information leads to the misclassification of noise points as target points, thereby diminishing the accuracy of target segmentation. Consequently, our experimental findings indicate that optimal results can be achieved when employing a sampling frame size of 32×32. In Fig. 8, the segmentation outcomes of our dataset are presented, featuring three distinct box sizes: 64×64, 32×32, and 16×16.

TABLE V
TESTING RESULTS OF DETECTION ON THE YOLO NETWORK

data	LSNR	mAP %	Recall %	precision %	F1
data19	4.469	85.89	87.00	78.14	0.8
data19+PSEnet	4.492/+0.023	97.13/+11.24	98.17/+11.17	77.70/-0.44	0.85/+0.05
data8	3.451	32.99	38.67	69.46	0.5
data8+PSEnet	3.469/+0.018	57.40/+24.01	62.83/+24.16	62.62/-6.84	0.63/+0.13
data11	0.886	22.30	29.00	21.97	0.25
data11+PSEnet	3.774/+2.888	49.82/+27.52	62.33/+33.33	36.38/+14.41	0.5/+0.25

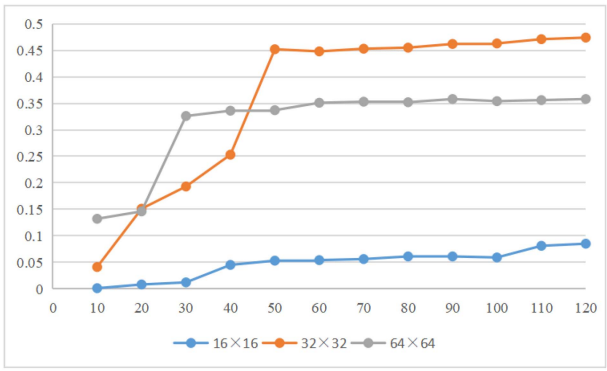

Fig. 9. F1 on sample times with three patch sizes.

We further analyze the test sets data8, data11, and data19. The figures provide a more intuitive understanding of the segmentation results. In the case of a simple scene (data19), when employing a sampling window size of 16 × 16, the segmentation result contains some noise. However, by increasing the sampling window to 32 × 32 or larger, the segmentation result improves. Moving on to the medium scene (data8), a sampling window size of 16 × 16 yields numerous noise points in the segmentation result while a window size of 64 × 64 fails to effectively separate the target points. Nonetheless, when the sampling window is set to 32 × 32, the segmentation results are accurate and nearly devoid of other noise. In the case of complex scenes (data11), setting the sampling window to 16 × 16 also leads to the presence of many noise points in the segmentation result. Although accurate segmentation of the targets can be achieved with a window size of 64 × 64, the brightness of the target points is weaker. Consequently, the segmentation result is not as satisfactory as when using a 32 × 32 sampling frame. The F1 score variations with respect to the number of samples are depicted in Fig. 9. It can be observed that after 50 sample times, the F1 scores for all sample sizes exhibit minimal changes. Hence, we choose 50 sample times as the optimal value.

To demonstrate that our method is not only effective in accomplishing background suppression, but also aids in downstream detection tasks, we investigated the impact of integrating our method prior to detection on the accuracy of detection results. The article uses the proposed enhancement algorithm to fuse the enhanced image with the original image to get a new input image and then put it into the detection network. We use YOLO [37] as the base model to test the detection accuracy of our enhanced algorithm because it can use a larger receptive field through the fusion of multiscale feature maps. It expresses the characteristics of small targets and is a very suitable base model for small-target detection. At the same time, we replace the measurement of the intersection ratio (IOU) of the predicted value and the real value with the 2-D Gaussian distribution distance [38]. This is because when the target size is very small and the predicted value has a slight deviation, it will lead to a sharp drop in the IOU, and the 2-D Gaussian distribution distance can well balance the error caused by the offset and enhance the robustness of the model.

The experimental findings presented in Table V demonstrate that the mAP is 85.89% for the simple scene, whereas it drops to 32.99% for the medium scene and 22.30% for the complex scene. These outcomes indicate a correlation between LSNR and the detection performance. Specifically, a higher LSNR leads to improved target detection results while a lower LSNR yields diminished performance. Further analysis of Table V reveals that our enhanced algorithm exhibits enhanced accuracy rates for simple scenarios, with recall and mAP both surpassing 97%. In the case of medium scenes, our algorithm significantly improves the recall, precision, and mAP, achieving 57.40% for medium scenes and 48.82% for complex scenes. These findings highlight the effectiveness of our enhanced algorithm in improving target detection performance across different scenes.

D. Compared With State-of-the-Art Methods

1) Numerical Evaluation: In order to accurately illustrate the effectiveness of the proposed method, we quantitatively evaluate it in comparison to the SOTA methods. Table VI presents the quantitative metrics for the compared methods on the "data8," "data11," and "data19." The traditional patch-based methods exhibit high precision and low recall. This is due to the fact that classical IR small-target detection methods focus more on detecting the location of the target and ignore the importance of complete segmentation of the target. These methods fail to detect the entire region of the target and can only detect a few pixels in the target region, which results in low recall for the comparison methods. In contrast, CNN-based methods seem to have more balanced precision and recall compared to patch-based methods. It can be seen that the proposed method is the best in terms of precision and F1 with relatively high improvement.

In "data8," our proposed method has the highest results on all measurements. All CNN-based approaches have very high precision and recall. Thanks to the PSEnet can fully utilize the local patch information and give a more detailed spatial information representation of the dim and small IR targets. While the other patch-based methods are sensitive to noise, which leads to low recall and precision in "data8." In "data11," the target has so weak energy that is submerged in the complicated background. The local contrast measurement is so easy that cannot calculate

TABLE VI
COMPARISON RESULTS OF IR DIM AND SMALL-TARGET BACKGROUND SUPPRESSION WITH SOTA METHODS

method	data8			data11			data19			Average		
	recall	precision	F1	recall	precision	F1	recall	precision	F1	recall	precision	F1
LCM	0.16	0.13	0.14	0.32	0.15	0.26	0.05	0.71	0.19	0.18	0.33	0.20
MPCM	0.2	0.07	0.1	0.11	0.16	0.18	0.07	0.21	0.14	0.13	0.15	0.14
AMWLCM	0.21	0.61	0.32	0.23	0.36	0.27	0.29	0.57	0.38	0.24	0.51	0.32
Top-Hat	0.14	0.05	0.08	0.23	0.29	0.44	0.01	0.69	0.17	0.13	0.34	0.23
Max-median	0.07	0.67	0.24	0.45	0.92	0.65	0.32	0.51	0.33	0.28	0.70	0.41
ALCNet	0.62	0.58	0.6	0.49	0.34	0.45	0.54	0.47	0.52	0.55	0.46	0.52
MDvsFAcGAN	0.82	0.88	0.83	0.69	0.87	0.75	0.71	0.65	0.69	0.74	0.80	0.76
IAANet	0.85	0.89	0.87	0.73	0.89	0.77	0.72	0.73	0.72	0.77	0.84	0.79
AGPCNet	0.89	0.91	0.9	**0.76**	0.88	0.78	0.74	0.77	0.75	0.80	0.85	0.81
OURS	**0.95**	**0.92**	**0.94**	0.75	**0.93**	**0.82**	**0.75**	**0.78**	**0.77**	**0.82**	**0.88**	**0.84**

The best results are bolded and the second best results are underlined.

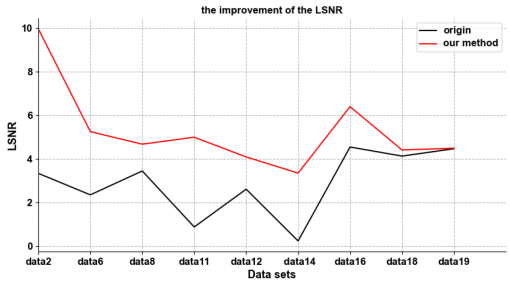

Fig. 10. Comparison between the origin LSNR and the enhanced LSNR.

the target area correctly. As a result, the F1 is lower than other data. But for the PSEnet can learn the spatial and structure information rather than the contrast information, the dim and small target can be found more easily than other CNN-based methods. In "data19," the bottom of the image has many noises such as houses and cars, which is similar to the target. Thus, the precision of "data19" is decreased. But, the PSEnet uses the multilayer features from the shallow layers and hollow layers, the targets can be easily segmented from the background noises. For all data in the dataset, the proposed methods achieve the best results in precision, recall, and F1. It can be concluded that the proposed method can increase the SNR of the target area and suppress the background.

2) Visual Evaluation: In comparison with CNN-based methods, because the PSEnet can utilize the local information of dim and small IR targets and improve the feature texture representation of the dim and small targets. The missing detection rates and false alarm rates in all sequences are decreased. According to the formula of the LSNR of the image, it can be seen that when the difference between the mean brightness of the target area and the background area is larger, the LSNR of the image will also be larger. From the discount graph analysis, the proposed method improves the LSNR for each image as shown in Fig. 10, especially for sequences with a low LSNR. From the analysis of the visualization results, the proposed method can get a background suppression map, which estimates the position of the target very accurately and suppresses the noise caused by the background.

As shown in Figs. 11–13, we selected three typical IR small-target scenes and compared the detection results of nine methods. In the figures, we use red boxes to mark the target position, blue boxes to mark the missing target area, and yellow boxes to highlight the imperceptible false alarms area. The detection results of our proposed method are shown in the right bottom corner.

As can be seen from the figure, the Tophat method exhibits sensitivity to noise, leading to a strong response to both noise and edge clutter in the background during detection. Similarly, the detection results of the LCM are unsatisfactory in low SNR scenes. Furthermore, although MPCM is capable of detecting targets, a significant amount of clutter remains in the background. This can be attributed to the simplistic background assumption of the HVS, making it challenging to distinguish between background and targets using global information alone when their characteristics are similar. Conventional patch-based methods often experience high rates of missed detections. In addition, when compared with the ground truth, it is evident that these methods only provide an approximate estimation of the target's location. By incorporating various constraints guided by prior knowledge, these methods impose strict limitations on the targets, resulting in small segmented target areas in the output. Moreover, the figures illustrate that these methods that heavily rely on models, assumptions, and parameters are not robust.

In summary, patch-based approaches usually have higher precision, but recall is very low. This is because there are many false alarms in the results of such methods. At the same time, they put more emphasis on detecting targets instead of accurately segmenting targets.

V. DISCUSSION

Our PSEnet pioneers a shift from the conventional global perspective to a detailed, localized focus, making the process of background suppression more straightforward and efficient. It implements a unique patch-based mechanism, focusing on

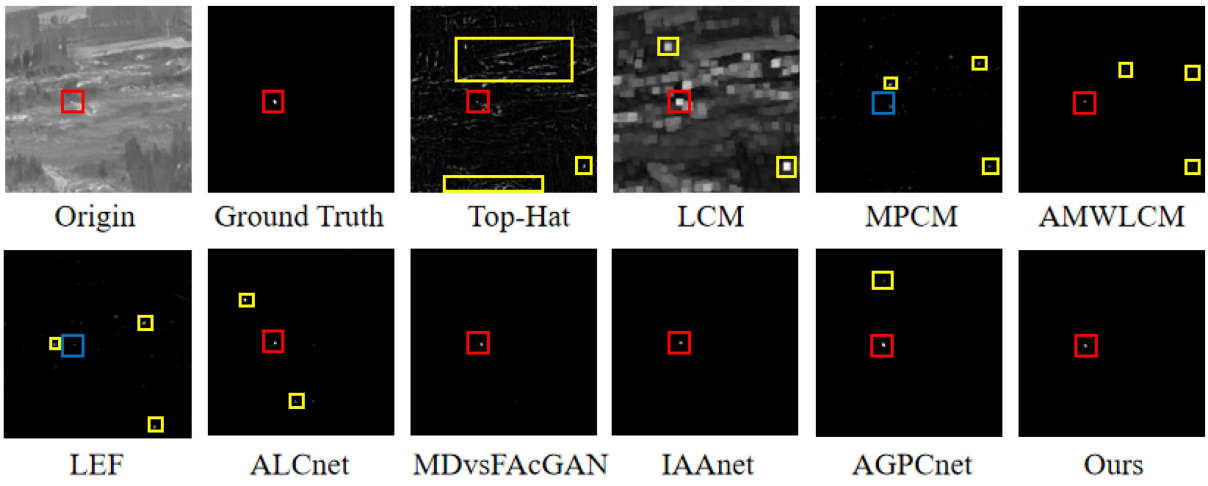

Fig. 11. Detection results of PSEnet and baseline methods on "data8."

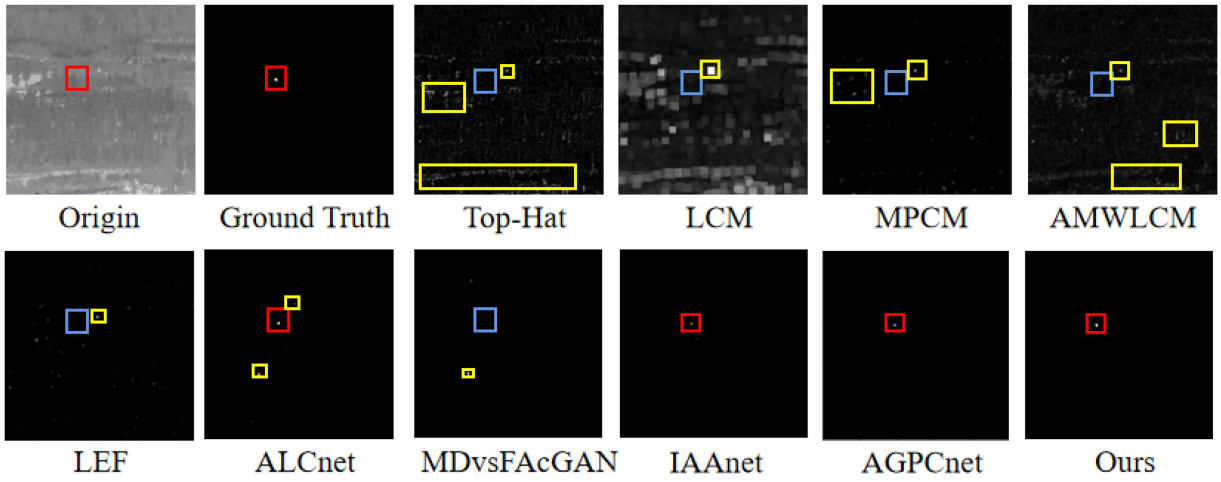

Fig. 12. Detection results of PSEnet and baseline methods on "data11."

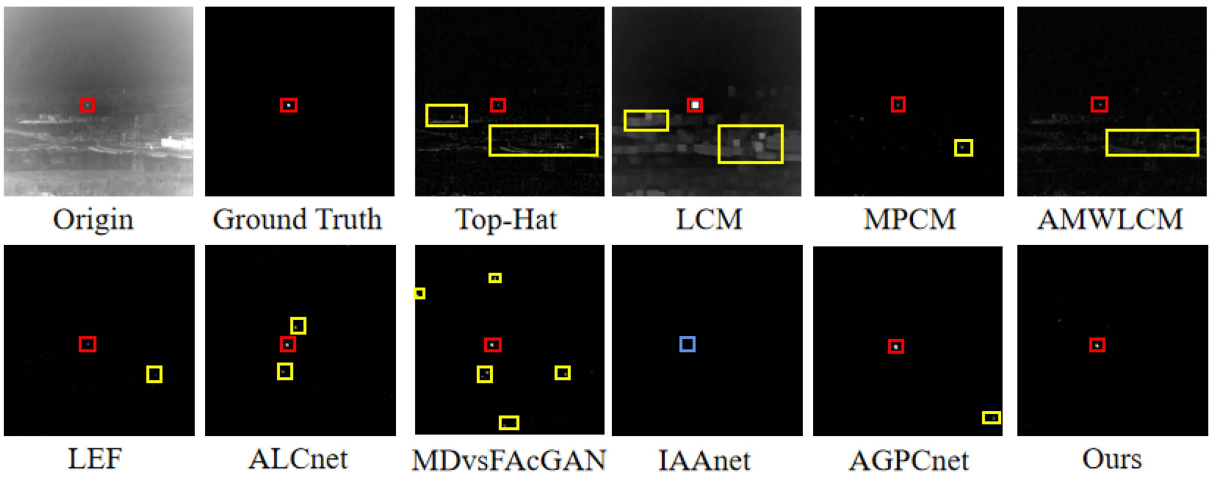

Fig. 13. Detection results of PSEnet and baseline methods on "data19."

individual patches within a comprehensive image, which simplifies computations while enhancing target attribute isolation. Furthermore, we incorporate multiscale dilated convolution and adaptive weight fusion within the PSE module to extract enriching local semantic information. In addition, our research presents an innovative sampling strategy within the SP module, introducing balance at the patch level and leading to an improved dataset distribution. The novel generator-discriminator dynamic in adversarial network training continually improves the quality of generated images.

Although the proposed method has retained and highlighted the features of dim and small targets, there are still some constraints in this article. The proposed method cannot use the global information and time information of the frames to train, so it cannot accurately segment the target completely obscured by clouds. In future research, the time sequence information between frames will be integrated into the background suppression network to help the model segment targets more accurately. Meanwhile, it will be an important task to study the global and local fusion methods to get a more robust feature representation of dim and small targets.

VI. Conclusion

In conclusion, we propose an innovative background suppression method in the realm of long-distance dim and small-target detection within IR imagery. This holistic approach is uniquely characterized by its framework, which consists of three critical constituents—1)the SP module, 2) the patch-based semantically enhanced GAN (PSEnet), and 3) the LTG module. The SP module enriches our training sample universe via localized sampling, thereby augmenting the representation of small-target features. This technique has proven instrumental in navigating the complexities of intricate background suppression. Build on this, PSEnet simplifies computations and facilitates the extraction of local semantic information, thus accelerating target identification and isolation processes. The LTG module coherently maps the local patches to the global image, thereby adhering to the algorithm's pursuit of background suppression and clear target distinction. Finally, through rigorous testing and experimentation, our novel approach has proven its mettle by delivering superior detection accuracy, particularly in low-SNR scenarios, and shall continue to redefine the boundaries of IR target detection innovation.

References

[1] 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, Mar. 2022, Art. no. 5002013.

[2] W. Haoxian, D. Heng, and Z. Zhiquan, "Review on dim small target detection technologies in infrared single frame images," *Laser Optoelectron. Prog.*, vol. 56, no. 8, 2019, Art. no. 080001.

[3] C. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.

[4] J. Liu, Z. He, Z. Chen, and L. Shao, "Tiny and dim infrared target detection based on weighted local contrast," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1780–1784, Nov. 2018.

[5] Y. Shi, Y. Wei, H. Yao, D. Pan, and G. Xiao, "High-boost-based multiscale local contrast measure for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 33–37, Jan. 2018.

[6] 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, Jul. 2021, Art. no. 5001015.

[7] Y. Li et al., "Infrared maritime dim small target detection based on spatiotemporal cues and directional morphological filtering," *Infrared Phys. Technol.*, vol. 115, 2021, Art. no. 103657.

[8] Y. Li, Y. Zhang, J.-G. Yu, Y. Tan, J. Tian, and J. Ma, "A novel spatio-temporal saliency approach for robust dim moving target detection from airborne infrared image sequences," *Inf. Sci.*, vol. 369, pp. 548–563, 2016.

[9] 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.

[10] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 950–959.

[11] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1752, Nov. 2022.

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

[13] 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.*, 2019, pp. 8509–8518.

[14] Z. Wang, C. Cao, and Y. Zhu, "Entropy and confidence-based undersampling boosting random forests for imbalanced problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5178–5191, Dec. 2020.

[15] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *Proc. 11th Int. Conf. Inf. Commun. Syst.*, 2020, pp. 243–248.

[16] B. Bosquet, D. Cores, L. Seidenari, V. M. Brea, M. Mucientes, and A. D. Bimbo, "A full data augmentation pipeline for small object detection based on generative adversarial networks," *Pattern Recognit.*, vol. 133, 2023, Art. no. 108998.

[17] 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.*, 2017, pp. 2980–2988.

[18] 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.

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

[20] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," in *Signal Data Process. Small Targets* Bellingham, WA, USA: Int. Soc. Opt. Photon., vol. 3809, 1999, pp. 74–83.

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

[22] X. Kong, C. Yang, S. Cao, C. Li, and Z. Peng, "Infrared small target detection via nonconvex tensor fibered rank approximation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2021, Art. no. 5000321.

[23] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, 2016.

[24] S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Comput. Vision, Graph., Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.

[25] 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.

[26] 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.

[27] T. Liu et al., "Nonconvex tensor low-rank approximation for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2021, Art. no. 5614718.

[28] 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.

[29] L. Zhang, L. Peng, T. Zhang, S. Cao, and Z. Peng, "Infrared small target detection via non-convex rank approximation minimization joint l2, 1 norm," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1821.
[30] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 382.
[31] Y. Dai, Y. Wu, Y. Song, and J. Guo, "Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values," *Infrared Phys. Technol.*, vol. 81, pp. 182–194, 2017.
[32] Z. Wang, Z. Fang, D. Li, H. Yang, and W. Du, "Semantic supplementary network with prior information for multi-label image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1848–1859, Apr. 2022.
[33] X. Kuang, X. Sui, Y. Liu, Q. Chen, and G. Gu, "Single infrared image enhancement using a deep convolutional neural network," *Neurocomputing*, vol. 332, pp. 119–128, 2019.
[34] X. Ying et al., "MoCoPnet: Exploring local motion and contrast priors for infrared small target super-resolution," 2022, *arXiv:2201.01014*.
[35] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1432.
[36] B. Hui, Z. Song, and H. Fan, "A dataset for infrared detection and tracking of dim-small aircraft targets under ground/air background," *China Sci. Data*, vol. 5, no. 3, pp. 291–302, 2020.
[37] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[38] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.

Yunfei Tong is currently working toward the Ph.D. degree in computer science with the East China University of Science and Technology, Shanghai, China.

Her current research interests include small-target detection, IR small-target tracking, and model lightweighting.

Yue Leng received the M.Eng. degree in computer application and technology from the Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, China, in 2023.

Her current research interests include small-target detection and machine learning.

Hai Yang received the B.Sc. degree in software engineering from Xi'an Jiaotong University, Xi'an, China, in 2008, and the Ph.D. degree in signal and information processing from the Chinese Academy of Sciences University, Beijing, China, in 2013.

He is currently a Research Associate Professor with the Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, China. At present, he has more than 10 papers published in many famous international journals, including *Bioinformatics*, *Nature Neuroscience*, etc. His research interests include artificial intelligence, machine learning, big data, and bioinformatics.

Zhe Wang (Associate Member, IEEE) received the B.Sc. degree in computer science and technology and the Ph.D. degree in computer application and technology from the Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 2003 and 2008, respectively.

He is currently a Full Professor with the Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, China. At present, he has more than 60 papers with the first or corresponding author published in some famous international journals including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING, *Pattern Recognition*, etc. His research interests include feature extraction, kernel-based methods, image processing, and pattern recognition.

Saisai Niu received the Ph.D. degree in aerospace manufacturing engineering from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2013.

He is currently a Senior Engineer with the Shanghai Aerospace Control Technology Institute, China Aerospace Science and Technology Corporation, Beijing, China. In 2019, he attended the space studies program (SSP2019) course with International Space University, France. In recent years, he has authored or coauthored more than 20 papers in *Sensor*, *Infrared Physics and Technology*, IEEE Xplore, IEEE ACCESS, and other peer-reviewed journals. He is undertaking multiple National Defense Key Research Projects and the National Basic Research Program. At present, his main research interests include IR target characteristics, IR scene modeling, semiphysical simulation, photoelectric detection guidance, artificial intelligence technology, and its application in IR precise guidance systems.

Huabao Long received the bachelor's degree from Northwestern Polytechnical University, Xi'an, China, in 1999, and the master's degree from Shanghai Jiao Tong University, Shanghai, China, in 2013.

He has participated in multiple development tasks of photoelectric detection systems, was the leader of multiple technological innovation projects, and is an academic and technical leader. He has authored or coauthored multiple SCI/EI and core journal papers, applied for more than 20 patents, and obtained 10 authorizations. At present, his main research interests include IR detection, target identification, and information processing technology.

Mr. Long won a third prize in the National Defense Science and Technology Progress Award.