Mitigate Target-Level Insensitivity of Infrared Small Target Detection via Posterior Distribution Modeling

Haoqing Li, Jinfu Yang, Yifei Xu, and Runshi Wang

***Abstract*—Infrared small target detection (IRSTD) aims to segment small targets from infrared clutter background. Existing methods mainly focus on discriminative approaches, i.e., a pixel-level front-background binary segmentation. Since infrared small targets are small and low signal-to-clutter ratio, empirical risk has few disturbances when a certain false alarm and missed detection exist, which seriously affect the further improvement of such methods. Motivated by the dense prediction generative methods, in this article, we compensate pixel-level discriminant with mask posterior distribution modeling. Specifically, we propose a diffusion model framework for IRSTD. This generative framework maximizes the posterior distribution of the small target mask to surmount the performance bottleneck associated with minimizing discriminative empirical risk. This transition from the discriminative paradigm to generative one enables us to bypass the target-level insensitivity. Furthermore, we design a low-frequency isolation in wavelet domain to suppress the interference of intrinsic infrared noise on the diffusion noise estimation. The low-frequency component of the infrared image in the wavelet domain is processed by a neural network, and the high-frequency component is utilized to restore the targets information, to estimate the residuals of the enhanced features. Experiments show that the proposed method achieves competitive performance gains over state-of-the-art methods on NUAA-SIRST, NUDT-SIRST, and IRSTD-1 k datasets.**

***Index Terms*—Deep learning, diffusion model, generative model, infrared small target detection (IRSTD).**

I. Introduction

INFRARED small target detection (IRSTD), a technique for finding small targets from infrared clutter background, provides many potential applications in remote sense, public security, and other fields with strong antidisturbance imaging capability [1], [2], [3]. However, infrared small targets tend to be smaller than 9×9 pixels, insufficiency of color and texture features, and consequently, submerged by a clutter background. These constraints make IRSTD challenging and prone to false alarm and missing detection. Conventional unlearnable methods, for instance, filtering-based [4], [5], local contrast-based [6], [7], and low rank-based methods [8], [55] are hindered by numerous hyperparameters. Concurrently, their pixel-level performance is relatively inferior, with high false alarm and missing detection. In addition, deep learning methods [9], [10], [11], [12] have achieved better pixel-level accuracy and lower false alarm and miss detection. Dai et al. [13] used discriminative deep learning earlier and made a pioneering breakthrough. Li et al. [14] proposed a dense nested attention network to incorporate and exploit contextual information. Zhang et al. [15] emphasized that shape matters and incorporated shape reconstruction into IRSTD. At present, the mainstream deep learning IRSTD methods are per-pixel discriminative learning, with intersection over union (IoU) [16] or binary cross entropy (BCE) loss as the cost function. For example, methodologies such as ACM [13] and DNANet [14] calculate the IoU between predicted masks and ground truth as empirical loss, whereas UIUNet [10] and ILNet [17] employ pixel-level BCE for optimizing model parameters. However, due to the minuscule size of infrared small targets, the ratio of target pixels to the overall infrared image is exceedingly small. Consequently, the cost function has no significant disturbances when a certain false alarm and missed detection exist, owing to the rare pixels in the targets. The inconspicuous error is manifested in the false alarm rate (the prevailing state-of-the-art methods typically fall within the order of 10^{-6}). Especially in the later training period with declined low learning rate, this insensitivity of the empirical risk seriously affects the further improvement of such methods.

How to avoid target-level insensitivity is the key to solving the above problem. The discriminantive loss functions only consider the local differences, which are not significant in IRSTD. How to obtain general rules for the existence of targets from a global perspective? Obtaining the posterior distribution of the target mask directly is one way. Inspired by Le et al. [18], denoising diffusion probabilistic models (DDPM) can model the conditional distribution of binary masks, providing a perspective for nonpixel-level discrimination in intensive prediction tasks.

In this article, the generative diffusion framework is employed to obtain the mask distribution directly, instead of pixel-level discrimination. The scattered false alarm pixels bring strong disturbances to the simple data pattern, but the KL divergence constraint on the mask posterior distribution can suppress such outliers. Therefore, given the conditions of the input infrared image, we maximize the posterior distribution of the small target mask to surmount the performance bottleneck associated with

Manuscript received 11 June 2024; revised 5 July 2024; accepted 13 July 2024. Date of publication 17 July 2024; date of current version 5 August 2024. This work was supported by the National Natural Science Foundation of China under Grant 61973009. *(Corresponding author: Jinfu Yang.)*

Haoqing Li, Yifei Xu, and Runshi Wang are with the Faculty of Information, Beijing University of Technology, Beijing 100124, China (e-mail: lihaoqing@emails.bjut.edu.cn; xuyifei@emails.bjut.edu.cn; wangrunshi@emails.bjut.edu.cn).

Jinfu Yang is with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China, and also with the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124, China (e-mail: jfyang@bjut.edu.cn).

Data is available on-line at https://github.com/Li-Haoqing/IRSTD-Diff.

Digital Object Identifier 10.1109/JSTARS.2024.3429491

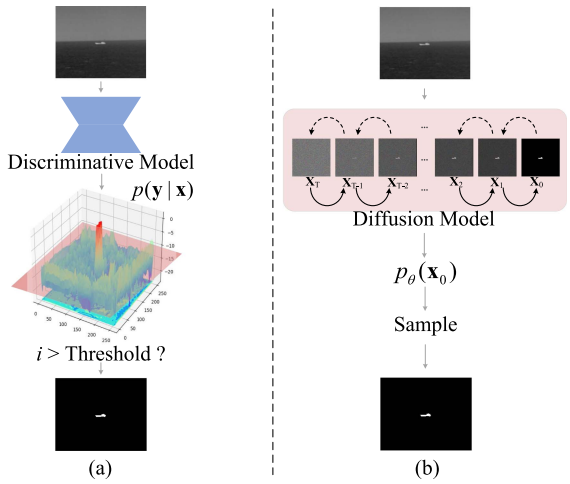

Fig. 1. Schematic comparison between (a) existing discriminative methods and (b) our generative method for IRSTD. Our approach obtains the posterior distribution of small target masks.

minimizing discriminative empirical risk. The final detection results are obtained by sampling from this distribution, as shown in Fig. 1.

However, there is inevitably a massive amount of noise in the infrared imaging process, including background and hardware noise. The global white noise under the influence of background radiation, as well as shot noise, thermal noise, and low-frequency 1 / f noise caused by hardware in the thermal imaging, leads to an extremely low signal-to-clutter ratio [19], [20]. These interferences will affect the noise estimation of the diffusion model. Several prophase image processing approaches [21], [22], [23], [24] used wavelet transform [25], i.e., transform-filter-inverse, to enhance the infrared image, which was beneficial to IRSTD and other tasks. Nevertheless, these complex approaches cannot be applied to end-to-end diffusion model training. Based on the basic concept that "the infrared background remains in the low-frequency component, while the targets with higher intensity are reflected in the high-frequency" [26], a low-frequency isolation module in the wavelet domain is designed. The low-frequency component of the infrared image in the wavelet domain is processed by a neural network, and the high-frequency component is utilized to restore the targets information. Finally, a residual of the enhanced features is estimated.

The contributions of this article can be summarized as follows.

1) We examine the back-propagation of IRSTD and highlight a potential issue in discriminative training, i.e., the target-level insensitivity.
2) A diffusion model framework for IRSTD is proposed. We generatively obtain the posterior distribution of small target masks, to surmount the performance bottleneck associated with minimizing discriminative empirical risk.
3) A low-frequency isolation module in the wavelet domain is designed, to reduce the influence of low-level infrared interference on diffusion noise estimation.
4) Qualitative and quantitative experiments are evaluated on three datasets and demonstrate better performance than the state-of-the-art discriminative methods.

II. Related Work

A. Infrared Small Target Detection

IRSTD is widely investigated using different approaches. These include unlearnable methods: filtering-based, local contrast-based, and low rank-based methods; and deep learning methods: ACM [13], DNA [14], ISNet [15], etc. The data-driven deep learning methods have better accuracy and convenient inference, without numerous manual hyperparameters. Dai et al. [13] first introduced the data-driven deep learning method into IRSTD and proposed an asymmetric contextual modulation module. Furthermore, they combined discriminative networks and conventional model-driven methods to utilize the infrared images and domain knowledge [9]. Zhang et al. [27] proposed an attention-guided pyramid context network and further improved the applicability of data-driven methods of IRSTD. Li et al. [14] proposed a dense nested attention network and a channel-spatial attention module for adaptive feature fusion and enhancement, to incorporate and exploit contextual information. Zhang et al. [15] indicated that shape matters for IRSTD and proposed Taylor finite difference (TFD)-inspired edge block aggregates to enhance the comprehensive edge information from different levels. Our previous work [17] found that the infrared small target lost information in the deep layers of the network, and proved that the low-level feature matters for recovering the disappeared target information. These data-driven deep learning methods are discriminative, which are prone to generate plenty of false alarms and miss detection in the process of pixel-level discrimination. This phenomenon caused by the target-level insensitivity during discriminative training. In this article, we surmount this problem with generative posterior distribution modeling.

B. Diffusion Model

Diffusion model [28], a prevailing and promising generative model at present, is superior to other types of generative models such as GANs [29], [30] and VAEs [31], [32]. The conditional diffusion probabilistic models [33], [34] use class embedding to guide the image generation process to a certain extent. These methods have led to some excellent large-scale applications, e.g., Imagen [35], ERNIE-ViLG [36], and stable diffusion [37], which have made significant contributions to AI-generated content.

The diffusion model has been applied to many generative tasks, such as inpainting [38], image translation [39], super-resolution [40], etc. Similarly, the discriminative application has also been explored. Xiang et al. [41] proved that the denoising diffusion autoencoders (DDAE) are unified self-supervised learners and the diffusion pretraining is a general approach for self-supervised generative and discriminative learning. Therefore, the generative pretraining paradigm supports the diffusion model as a feasible discriminative learning method. Baranchuk et al. [42] demonstrate that diffusion models can serve as an instrument for dense prediction tasks, especially in the setup when labeled data are scarce. Simultaneously, Amit et al. [43] used the diffusion model for image semantic segmentation, Wu et al. [44], [45] used it to address the medical image segmentation problem,

and Ma et al. [46] demonstrated the superiority of adapting diffusion for unsupervised object discovery. These works, as the explorers applied the diffusion model to discriminative tasks, have proved their feasibility and excellent performance.

Additionally, Chen et al. [47] regarded semantic segmentation as an image-conditioned mask generation, which is replacing the conventional per-pixel discriminative learning with a latent prior learning process. Le et al. [18] modeled the underlying conditional distribution of a binary mask, which was conditioned on an object region and K-shot information. These methods provided a novel approach to pixel-level discriminative tasks such as IRSTD.

In this work, we replace pixel-level discriminant with mask posterior distribution modeling. A pertinence conditional diffusion model is employed to achieve this purpose.

III. Method

A. Motivation

During the experiments of IRSTD, we found that the empirical risk becomes quite low when the neural network is trained to the later epochs. But there are still some target-level errors, and these false alarm and miss detection are unable to be eliminated. What causes this phenomenon?

We conducted extensive research on the loss function and back-propagation of the IRSTD model [10], [13], [14], [17]. Our research reveals that the IoU loss demonstrates insensitivity to false alarms, but sensitivity to miss detection. BCE loss exhibits insensitivity to all the target-level errors. IoU loss is computed as the average within a minibatch

$$L_{\text{IoU}} = 1 - \frac{1}{N}\sum_{i}^{N}\frac{\text{TP}}{T + P - \text{TP}} \tag{1}$$

where N is the batch size, T, P, and TP denote the predicted, ground truth, and true positive pixels, respectively. For instance, in a batch of size 8, if there are nine false alarm pixels, the loss is a mere 0.0103, whereas the loss for a miss detection in a single sample amounts to 0.125 (assuming a resolution of 256 $\times$ 256). In addition, the BCE loss is more insensitive, yielding a loss of only 8×10^{-5} for the same false alarm scenario. Discriminative empirical risk is progressively insensitive as the resolution increases.

The insensitivity exerts a significant influence on the parameter updates, allowing these error phenomena unpunished. Illustrated in Fig. 2, the minimal empirical risk results in all pertinent partial derivatives $\frac{\partial L}{\partial w_i}(i = 0, 1, \ldots, n)$ exceedingly diminutive. Coupled with the learning rate decaying to a small value in the later epochs, the optimal parameters are struggling to be obtained. Our methodology amalgamates generative paradigm to address this performance bottleneck.

B. Background

Diffusion model [28], a type of generative latent variable model, includes *forward* and *reverse process*. The forward process adds Gaussian noise to the original image $\mathbf{x_0}$ according to a given variance table $[\beta_1, \ldots, \beta_T]$

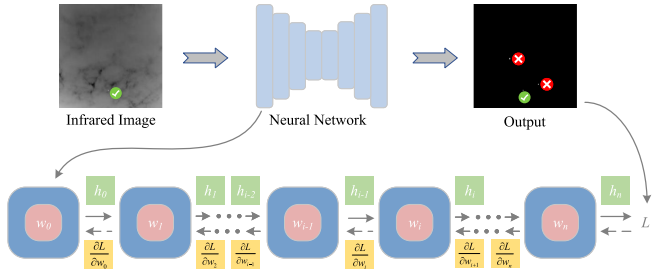

Fig. 2. Sketch map of the back-propagation process: The arrows to the right represent the forward-propagation, while the arrows to the left represent the back-propagation. For ease of representation, the bias and activation layers are omitted. w_i and h_i are the parameter and output of the ith layer, respectively. All these partial derivatives in the IRSTD networks are exceedingly diminutive, but false alarms exist.

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{2}$$

The Markov process lasts T steps until the approximate standard Gaussian noise $\mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ is obtained

$$q(\mathbf{x}_{1:T}|\mathbf{x}_{t-1}) := \sum_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{3}$$

where $[\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots \mathbf{x}_T]$ is the implicit variable of the intermediate forward process, which can be obtained directly

$$q(\mathbf{x}_t|\mathbf{x}_0) := \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{4}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

The reverse process starts with a standard Gaussian noise $\mathbf{x}_T$, completes by T steps learnable Gaussian transformation, and obtains the data distribution $p_\theta(\mathbf{x}_0)$ finally, which can be formulated as

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{5}$$

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{6}$$

$$p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T})d\mathbf{x}_{1:T} \tag{7}$$

where $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t))$ is a learnable parameterized expectation, and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ is a parameterized noise. An unlearnable variance $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ with the configuration of [28] is utilized.

Training is to optimize the variable boundary of negative log-likelihood $\mathbb{E}[-\log p_\theta(\mathbf{x}_0)]$, and minimize the KL divergence between $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ and the forward process posterior $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, which can be simplified as

$$L = \sum_{t>1} D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \tag{8}$$

where $D_{\text{KL}}(.)$ represents the KL divergence

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}) \tag{9}$$

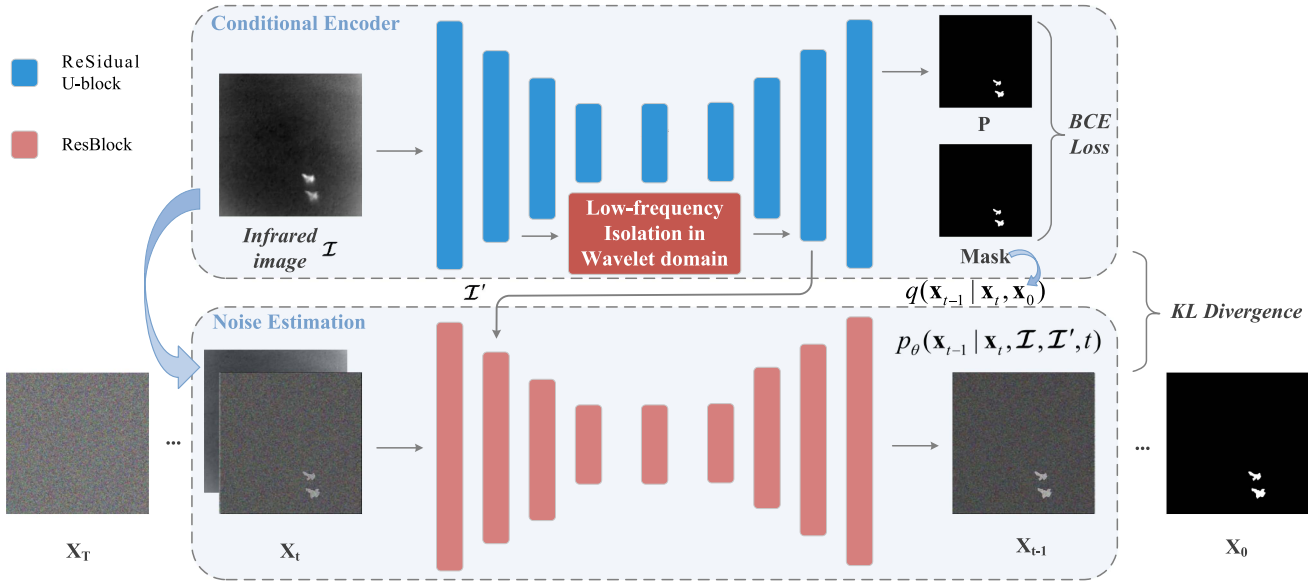

Fig. 3. Overview of the proposed diffusion framework for IRSTD. Each colored rectangle represents a corresponding block, and the rectangle's size reflects the latent space level of the blocks. The small gray arrows are in the same direction as the computational graph, while the large arrows are indicators.

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \tag{10}$$

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \tag{11}$$

If a neural network is used to predict noise $\boldsymbol{\epsilon}_\theta$, the Formula 8 can be equivalent to optimizing the conditional expectation

$$L' = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||^2] \tag{12}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I})$ is a standard Gaussian noise of forward process posteriors.

C. Overall Architecture

A diffusion-based approach, IRSTD-Diff, is proposed to obtain the underlying conditional mask distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathcal{I},t)$ for noised mask $\mathbf{x}_t$, conditioned on the infrared image $\mathcal{I}$, and latent space encoding $\mathcal{I}'$. Then, the posterior distribution of the original mask is obtained by the Markov process of Formula 6 and 7. The conditional distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathcal{I},t)$ can be formulated as

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t,\mathcal{I},\mathcal{I}',t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\mathcal{I},\mathcal{I}',t)\right) \tag{13}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathcal{I},\mathcal{I}',t) = \mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_\theta(\mathbf{x}_t,\mathcal{I},\mathcal{I}',t),\sqrt{\beta_t}\mathbf{I}) \tag{14}$$

where $\boldsymbol{\mu}_\theta$ is the parameterized expectation and $\boldsymbol{\epsilon}_\theta(.)$ is a noise estimation neural network.

D. Conditional Encoder

As shown in Fig. 3, the noise estimation network $\boldsymbol{\epsilon}_\theta(.)$ is based on the original U-Net [48] architecture of the DDPMs [28], [49] and combined with the conditional encoder (CE) network. CE, composed of residual U-block [50], is used to achieve the latent space perceptual compression of the infrared image $\mathcal{I}$, and capture the semantic information of small targets. Residual U-block is a network with sub-UNets embedded in the UNet architecture. The nested U-shaped structure enables the network to capture abundant local and global information from shallow and deep layers, which is applicable to all resolutions. Previous studies [10], [17] have demonstrated that this multilevel nested residual U-block as the backbone is more suitable for feature extraction of small infrared targets. Due to the high-level semantic information deficiency of small infrared targets, excessive latent space compression will remove the high-frequency target information. Therefore, we use the appropriate low-level latent space as the latent space embedding $\mathcal{I}'$. $\boldsymbol{\epsilon}_\theta(.)$ can be formulated as

$$\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\mathcal{I},\mathcal{I}',t) = U_{\text{de}}(U_{\text{en}}(\mathbf{x}_t,t) + f(\mathcal{I}),t) \tag{15}$$

where U_{de} and U_{en} denote U-Net encoder and decoder, respectively, and $f(.)$ is the CE network.

The intricate architecture of our CE network is delineated in Table II. This encoder employs a nested subencoding and decoding structure, comprising six encoders and five decoders internally. The subencoder is employed to achieve the latent space perceptual compression of infrared images. Throughout this process, the low-level information of the infrared image is compressed, and the semantic structure is preserved for the target reconstruction and restoration. This restoration reflects the necessity of the subdecoder's existence, i.e., BCE discriminative training. Our generative architecture is not to completely abandon the discriminative paradigm, which is crucial for pixel-level IRSTD. In the absence of joint discriminative training,

the sampling of the posterior distribution yield superior target-level performance, but struggle to achieve state-of-the-art pixel-level performance. This aspect will be thoroughly discussed in Section V.

Overall loss: To surmount the performance bottleneck of pixel-level discrimination, i.e., reduce false alarm and miss detection while maintaining pixel-level performance, we minimize the BCE-loss and the KL divergence between $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{I}}', t)$ and posterior $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ of the forward process

$$L = L' - \lambda[\mathbf{x}_0 \log(\mathbf{P}) + (1 - \mathbf{x}_0)\log(1 - \mathbf{P})] \quad (16)$$

where $\mathbf{P}$ is the estimated result of the CE, λ is a factor to balance an optimal learning rate of two parts.

E. Low-Frequency Isolation in the Wavelet Domain

During the imaging process, infrared images inevitably exhibit substantial noise, leading to an exceedingly low signal-to-noise ratio. It is inevitable to introduce unnecessary interference when using low-level latent encoding as $\boldsymbol{\mathcal{I}}'$, owing to the particularity of small infrared targets. Therefore, an LIW is designed, to reduce the low-level disturbance of the CE in diffusion noise estimation. The essence of infrared imaging is that the infrared noise is global. The background radiation predominantly consists of low-frequency components, whereas the higher-intensity targets are evident in the high-frequency components [26]. This should be considered from two aspects.

First, the wavelet transform demonstrates favorable localization properties in both the time and frequency domains, as well as multiscale features, facilitating image processing across various scales. The k-level 2-D Haar discrete wavelet transform (*HDWT*) [25] is applied to the original infrared features, as shown in Fig. 4 and Formula 17, to obtain the low-frequency approximation component $\mathbf{L}^k$ and the high-frequency vertical, horizontal, and diagonal components $\mathbf{H}_v^k, \mathbf{H}_h^k, \mathbf{H}_d^k$ $(k = 1, 2, \ldots K)$. This k-level 2-D *HDWT* of partial features is shown in Figs. 4 and 5. A substantial presence of background noise interference is evident in the low-frequency component, which is processed by the following neural network. However, mere low-frequency information cannot guarantee that high-frequency target features still exist after mapping. Figs. 4 and 5 demonstrate that the background features are scarce in high-frequency components, and the target features are more significant. Therefore, the high-frequency component is applied to the recovery of high-frequency target information. To sum up, the low-frequency approximate component $\mathbf{L}^k$ $(k = 1, 2, \ldots K)$ is utilized for residual mapping, as shown in Formulas 19 and 20. Then, 2-D Haar inverse discrete wavelet transform (HIDWT) based on the original high-frequency components $\mathbf{H}_v^k, \mathbf{H}_h^k, \mathbf{H}_d^k$ $(k = 1, 2, \ldots K)$ is applied to restore the infrared small target information, and initiated with the K-level low-frequency approximate component $\tilde{\mathbf{L}}^K = \mathbf{f}_i^K$, as shown in Formula 18.

Second, Swin Block [51] is employed to process the global disturbances of infrared features. Transformers [52] are marvelous at capturing the semantic structure of images and have strong global perception. It is applied for global semantic compression in the wavelet domain, to isolate low-frequency disturbances while preserving the semantic structure of infrared features

$$\mathbf{L}^k, \mathbf{H}_v^k, \mathbf{H}_h^k, \mathbf{H}_d^k = HDWT(\mathbf{f}_i^{k-1}) \quad (17)$$

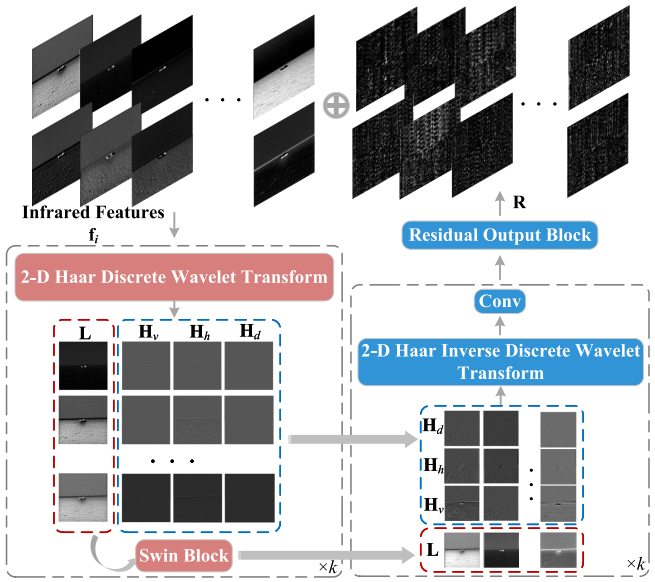

Fig. 4. Low-frequency isolation in the wavelet domain. L/H represent the visualization of high/low-frequency components in the wavelet domain. R is the visualization of the estimated residuals.

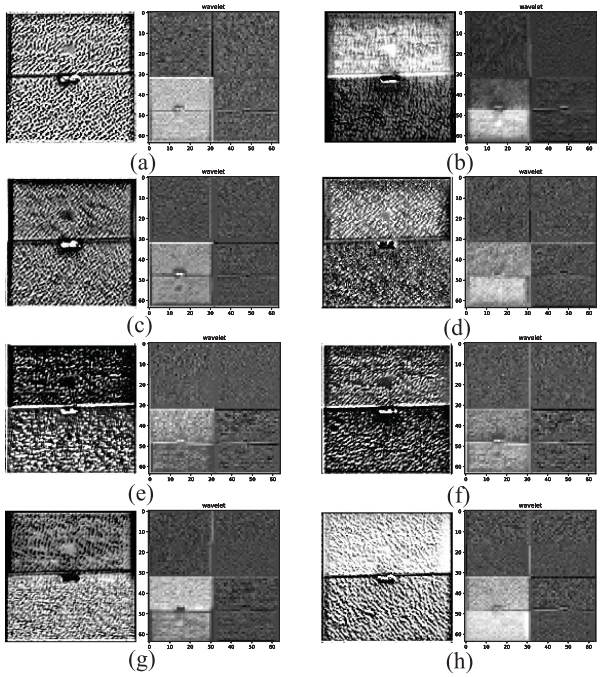

Fig. 5. One-level 2-D *HDWT* [25] applied to the infrared features. The left figures are CE output features, and the right figures are corresponding wavelet domain visualizations. The bottom left, bottom right, top left, and top right subfigures are low-frequency approximation and high-frequency horizontal, vertical, and diagonal components, respectively. These wavelet domain components are used as part of LIW.

$$\mathbf{f}_i^k = S(\mathbf{L}^k)$$
$$k = 1, 2, \ldots, K-1 \tag{18}$$

where *HDWT*(.) represents the 2-D *HDWT*, $\mathbf{f}_i^k$ is the *i*th stage features under *k*-level transform, and $S(.)$ denotes the Swin Block.

To obtain the optimal features $\hat{\mathbf{f}}$, it is easier to optimize the residual mapping $\mathbf{f} = \hat{\mathbf{f}} + \mathbf{R}$ when the original mapping $\mathbf{f} \to \hat{\mathbf{f}}$ is an identitylike mapping [53], [54]. Finally, as shown in Fig. 4, LIW estimates a residual $\mathbf{R}$ to isolate low-frequency disturbances

$$\tilde{\mathbf{L}}^{k-1} = Conv(HIDWT(\tilde{\mathbf{L}}^k, \mathbf{H}_v^k, \mathbf{H}_h^k, \mathbf{H}_d^k))$$
$$k = K, K-1, \ldots, 2 \tag{19}$$
$$\mathbf{R} = RO(\tilde{\mathbf{L}}^1) \tag{20}$$

where *HIDWT*(.) represents the 2-D *HIDWT*, $Conv(.)$ is a convolutional module, $RO(.)$ represents the residual output block, and $\tilde{\mathbf{L}}^k$ is *k*-level inversed low-frequency approximate component.

IV. Experiments

A. Setup

1) Datasets: Our experiments are conducted on NUAA-SIRST [13], IRSTD-1k [15], and NUDT-SIRST [14] datasets. The NUAA-SIRST dataset has 341 training data and 86 testing data. About 90% of the infrared images contain one target, and the rest of the samples contain multiple targets. The IRSTD-1 k dataset has 800 training data and 201 testing data. It contains different types of small targets, such as unmanned aerial vehicles, animals, ships and vehicles, which capture infrared images at a long imaging distance in multiple locations. The dataset covers many scenarios, including oceans, rivers, fields, mountain areas, cities and clouds, and has severe background clutter noise. The NUAA-SIRST and IRSTD-1 k datasets are the real infrared image data collected in the real scenario, and are manually labeled with pixel-level mask. Different from the above two datasets, all infrared images in the NUDT-SIRST dataset are composed of infrared background and various simulated infrared small targets. The dataset contains a variety of background scenarios, such as cities, fields, oceans and clouds. For each scenario, the resolution of all infrared images is 256 × 256. The NUDT-SIRST dataset has 1062 training data and 265 testing data. These datasets has a variety of target types, sizes, and clutter backgrounds, which can reliably evaluate the performance of the deep neural network-based methods. We use the same division setting as original paper [13], [14], [15].

2) Evaluation Metrics: a) Intersection over Union: IoU, a common pixel-level evaluation metric of IRSTD, is the ratio of the intersection and union region between the predictions and the ground truth. IoU is calculated on the whole dataset and the correctness of each pixel has a great impact

$$\text{IoU} = \frac{A_i}{A_u} = \frac{\sum_{i=1}^n \text{TP}_i}{\sum_{i=1}^n T_i + P_i - \text{TP}_i} \tag{21}$$

where A_i and A_u are the intersection and union, respectively. T denotes the pixels predicted as the targets. P denotes the pixels of the ground truth targets. TP is the true positive pixels. n represent the number of infrared images in the test set.

b) Probability of detection (P_d): P_d evaluates the ratio of the true targets detected by the model to all the ground truth targets. It is a target-level evaluation metric. A detected target is considered as TP if the centroid derivation is less than 3 [14]

$$P_d = \frac{TP_{\text{sum}}}{T_{\text{sum}}} \tag{22}$$

where TP_{sum} represents the number of detected targets and T_{sum} represents all the targets in the test set.

c) False-alarm rate (Fa): A target-level evaluation metric, to measure the ratio of predicted target pixels that do not match the ground truth

$$F_a = \frac{\sum_{i=1}^n \text{FP}_i}{\text{ALL}} \tag{23}$$

where FP denotes the number of false alarm pixels, and ALL is all pixels in the image.

3) Implementation Details: AdamW [58] is utilized as the optimizer. Weight decay is 0 and the learning rate is 0.0001, with a linear learning rate strategy. As the configuration of the diffusion model shown in Table I, an inferior configuration is used to verify the feasibility, limited by the computing hardware. The implementation of a more generalized diffusion model configuration may yield additional enhancements in performance, but it necessitates elevated standards for the computing hardware. The data pattern of the infrared small target masks is oversimplified compared with standard visual images, with weak discreteness. Accordingly, the *step* is set to 100. It can be reduced to an smaller value (e.g., 30) to diminish sampling time and enhance efficiency. The input images are resized to 256 × 256 during training and testing. The IRSTD-Diff is trained for 80 000 steps.

4) Comparative Discriminative Methods: For the factors outside the model, such as learning rate strategy, data augmentation, weight initialization, training strategy, etc., are unified to make a fair comparison [13]. All the methods are trained 400 epochs with a batch size of 8.

5) Comparative Generative Methods: Use the unified configuration as Table I. Select eight different steps for testing and choose the best result.

All methods are implemented based on Pytorch 1.10.0 and conducted on an NVIDIA GeForce RTX 3080 Ti.

A detailed configuration of the CE is shown in Table II. All experiments in this article were completed with this configuration.

B. Quantitative Comparisons

The IRSTD-Diff is compared with 13 SOTA methods including unlearnable methods: WSLCM [7], TLLCM [6], MSLSTIPT [55], PSTNN [8], IPI [56]; discriminative methods: ACM [13], AGPCNet [27], DNANet [14], ISNet [15], UIUNet [10]; and generative methods: SigDiff [43], MedSegDiff [45], CIMD [57].

TABLE I
UNIFIED CONFIGURATION OF THE DIFFUSION MODELS

Image Size	Batch Size	Diffusion Step	Training Step	UNet Channels	Resblocks	Heads	Attention Resolutions	Channel Multiple	Input Channels	Output Channels
256	4	100	80000	64	2	4	'16'	(1, 1, 2, 2, 4, 4)	3	1

TABLE II
DETAILED CONFIGURATION OF THE CE; *IN_C*, *MID_C*, AND *OUT_C* REPRESENT THE INPUT, MIDDLE, AND OUTPUT CHANNELS, RESPECTIVELY; *UP–DOWN* INDICATES WHETHER TO USE UP/DOWN-SAMPLING

Modules	*In_C*	*Mid_C*	*Out_C*	*Up-Down*
Encoder1	3	16	64	True
Encoder2	64	16	64	True
Encoder3	64	32	128	True
Encoder4	128	32	256	True
Encoder5	256	32	256	False
Encoder6	256	64	256	False
Decoder5	512	64	256	False
Decoder4	512	32	128	True
Decoder3	256	32	64	True
Decoder2	128	16	64	True
Decoder1	128	16	64	True

As shown in Table III, nonlearning methods exhibit significant disadvantages compared to deep learning-based approaches. On the NUDT-SIRST dataset, the pixel-level accuracy (IoU) shows a disparity of −74.38%, and the target-level accuracy has differences of 40.15×10^{-6} Fa and −24.17 % Pd. At present, nonlearning paradigms have no advantage compared to the data-driven deep learning paradigms.

Our IRSTD-Diff attains state-of-the-art performance on three datasets. As shown in Table III, our IRSTD-Diff is excellent in the target-level performance compared with other discriminative methods. It has established substantial superiority through the comparison of false alarm rate. In contrast to DNANet, which excels in pixel-level performance, our approach results in reductions of 6.92×10^{-6}, 3.18×10^{-6}, and 1.64×10^{-6} on the three datasets. As for UIUNet, our method has advantages of 7.37×10^{-6}, 4.69×10^{-6}, and 1.84×10^{-6} on the three datasets, respectively. Regarding the probability of detection, our IRSTD-Diff demonstrates comparable performance to UIUNet on the NUAA-SIRST and NUTD-SIRST datasets. On the IRSTD-1 k dataset, IRSTD-Diff surpasses the UIUNet by 0.34 %. Our method is comprehensively superior to other discriminative models.

The aforementioned observations serve to validate the assumptions articulated in the Section I. At the later period of discriminative training, it is difficult to fluctuate the experience risk, owing to a small proportion of false alarm and missed detection pixels. It is even more demanding to punish this phenomenon because of the decayed learning rate. Our approach compensates for this issue from the perspective of generative learning, i.e., modeling the mask posterior distribution. The scattered false alarm pixels bring strong disturbances to the simple data pattern. Therefore, KL divergence constraint on the mask posterior distribution can suppress such outliers. Therefore, IRSTD-Diff has low false alarm and missed detection, and competitive pixel-level performance. Five of the experimental outcomes outperformed the state-of-the-art methods in the six comparison experiments conducted on the three datasets. Besides, IRSTD-Diff has competitive pixel-level performance. With the exception of a slightly 1.21% lower IoU on the NUAA-SIRST dataset compared to UIUNet, the IRSTD-Diff achieves the state-of-the-art performance on the other two datasets, with improvements of 1.24% IoU compared to ISNet and 0.16% IoU to UIUNet.

Similarly, our IRSTD-Diff has achieved comprehensive advantages compared with other diffusion-based methods. The methods we compared with is the latest diffusion-based semantic segmentation or medical image segmentation approaches [43], [45], [57]. These excellent diffusion-based methods have made a pioneering contribution to the discriminative DDPM application. However, it is difficult to directly apply to IRSTD owing to the absence of pertinence design, resulting in inferior pixel-level and Pd performance. However, we observe a counterintuitive anomaly: these nonpertinence designed, diffusion-based dense prediction methods attain competitive false alarm rate. CIMD [57], a diffusion-based ambiguous medical image segmentation network, produces multiple plausible outputs by learning a distribution over group insights. We transfer it to the IRSTD task without modification. The false alarm rate of CIMD surpassed state-of-the-art discriminative methods by 5.15×10^{-6}, 3.65×10^{-6}, and 1.44×10^{-6} Fa on three datasets, respectively. The same phenomenon is also reflected in SigDiff [43], a cityscape semantic segmentation approach. SigDiff also delivered performance approaching the state-of-the-art methods, with the 4.66×10^{-6}, 1.82×10^{-6}, and -1.80×10^{-6} Fa advantages on three datasets, respectively. This atypical phenomenon further corroborates the hypothesis proposed in Section I, i.e., generative diffusion-based approaches effectively restrain minimal false alarm pixels that would not significantly disturb empirical risk.

In this work, the diffusion model for IRSTD is pertinence improved, which makes the diffusion model truly applicable to the issue of IRSTD. Our IRSTD-Diff demonstrates superior performance at the target-level, particularly in Pd, with improvements of 1.83%, 5.82%, and 0.81% compared to the optimal CIMD. Our pertinence enhancements, specifically designed to the intrinsic characteristics of infrared small targets, significantly heighten pixel-level accuracy, yielding 3.09%, 3.39%, and 2.99% IoU increasements on the three datasets, respectively. These results establish the effectiveness of our generative diffusion-based approach and emphasize the applicability for IRSTD.

The receiver operating characteristic (ROC) curves shown in Fig. 7 demonstrate the superiority and stability of our IRSTD-Diff. We quantitatively compare our IRSTD-Diff with other SOTA methods in area under curve (AUC), as shown in Table IV.

TABLE III
COMPARISONS WITH SOTA METHODS ON NUAA-SIRST, IRSTD-1 K, AND NUDT-SIRST DATASETS IN IOU(%), PD(%), AND $Fa(10^{-6})$

Methods	Description	NUAA-SIRST			IRSTD-1k			NUDT-SIRST		
		IoU	$F_a \downarrow$	P_d	IoU	$F_a \downarrow$	P_d	IoU	$F_a \downarrow$	P_d
WSLCM [7]	Local Contrast Based	1.16	5446	77.95	3.45	6619	72.44	2.28	1309	56.82
TLLCM [6]	Local Contrast Based	11.03	7.27	79.47	5.36	4.93	63.97	7.06	46.12	62.01
MSLSTIPI [55]	Local Rank Based	10.30	1131	82.13	10.37	3707	57.05	8.34	888.1	47.40
PSTNN [8]	Local Rank Based	22.40	29.11	77.95	24.57	35.26	71.99	14.85	44.17	66.13
IPI [56]	Local Rank Based	25.67	11.47	85.55	27.92	16.18	81.37	17.76	41.23	74.49
ACM [13]	Discriminative CNN Based	63.06	4.04	94.50	55.38	22.21	83.22	65.94	5.60	94.09
AGPC [27]	Discriminative CNN Based	69.08	4.48	98.17	61.00	12.96	88.36	84.49	2.69	97.58
ISNet [15]	Discriminative CNN Based	71.12	7.14	98.17	64.47	11.52	92.12	81.35	4.66	96.77
DNANet [14]	Discriminative CNN Based	69.85	8.38	98.17	64.44	8.98	92.12	89.57	2.72	98.12
UIUNet [10]	Discriminative CNN Based	**75.30**	8.83	**99.08**	63.68	10.49	92.81	91.98	2.92	**98.66**
SigDiff [43]	Diffusion Based	66.00	4.17	94.50	47.41	8.67	81.16	75.53	4.72	94.89
MedSegDiff [45]	Diffusion Based	64.19	16.80	97.25	57.65	21.94	81.16	78.27	10.15	96.50
CIMD [57]	Diffusion Based	71.00	3.68	97.25	62.32	6.84	87.33	89.15	1.48	97.31
IRSTD-Diff(ours)	Diffusion Based	74.09	**1.46**	**99.08**	**65.71**	**5.80**	**93.15**	**92.14**	**1.08**	98.12

The best result are in bold.

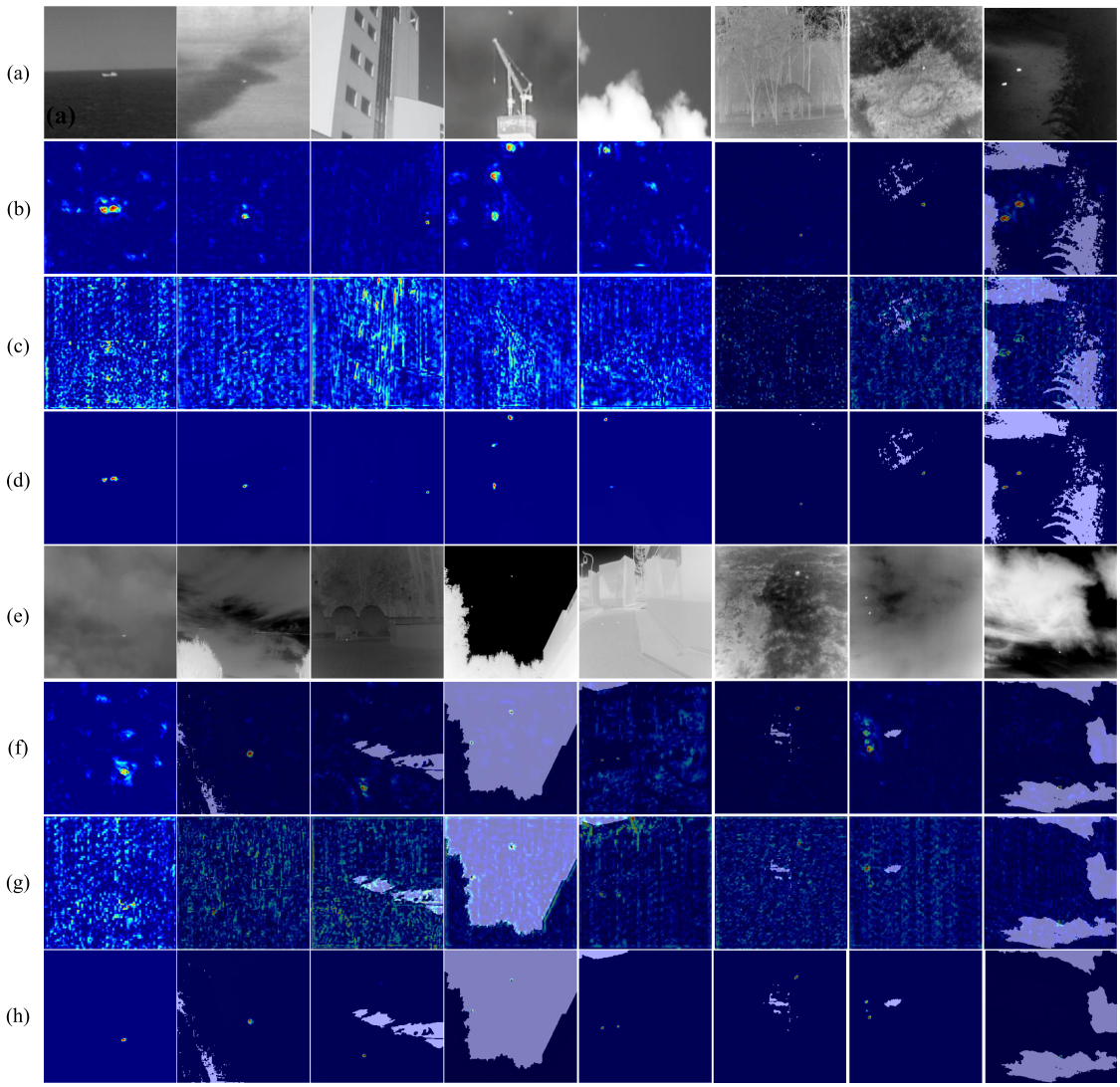

Fig. 6. Class activation mapping (CAM) comparison between (d), (h) our IRSTD-Diff and (b), (f) our model without LIW. (a) and (e) Input infrared images. LIW conspicuously isolates the background radiation and obtains a cleaner infrared CE. (c) and (g) Residuals $\mathbf{R}$ estimated by LIW for isolating low-frequency disturbances.

TABLE IV
AUC COMPARISION WITH OTHER METHODS

Methods	IPI	ACM	AGPC	DNANet	ISNet	UIUNet	SigDiff	MedSegDiff	CIMD	IRSTD-Diff
AUC	0.4545	0.6924	0.7019	0.7534	0.7692	0.8303	0.6727	0.6680	0.7828	0.7962

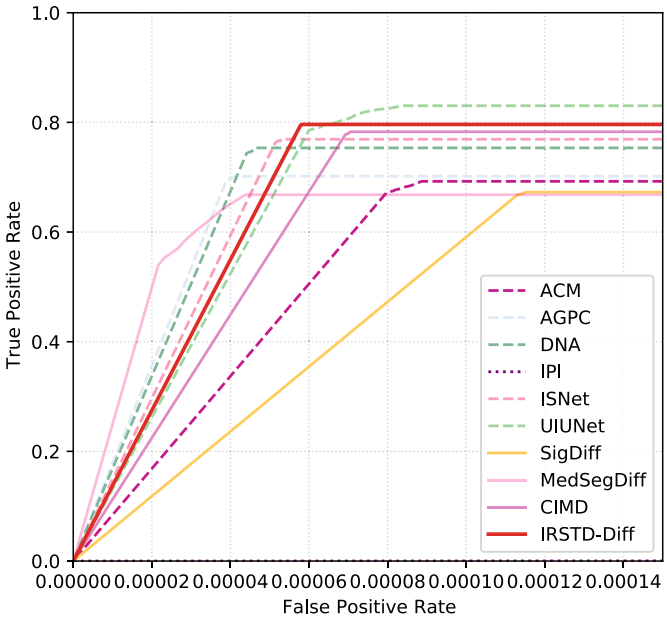

Fig. 7. ROC curves performance on the IRSTD-1 k dataset. The performance of our IRSTD-Diff is stable and better than other diffusion-based methods. It also has competitiveness compared with the discriminant methods.

TABLE V
EFFECTS OF THE k IN LIW; THE TRAINING STEP IS SET TO 100, AND WE STOCHASTICALLY CHOOSE A STATIONARY POSTERIOR PROBABILITY (60 000 STEPS) FOR INFERENCE

k	4	3	2	1
$IoU(\%)$	62.06	62.17	64.98	61.80
$F_a(10^{-6})\downarrow$	8.17	6.37	7.42	6.00
$P_d(\%)$	91.10	91.10	93.15	91.44

TABLE VI
EFFECTS OF THE DIFFUSION TRAINING STEPS; SMALL STEPS CAN ACHIEVE STUPENDOUS PERFORMANCE; LIMITED BY β, STEP CANNOT BE LESS THAN 20; *NW* MEANS NOT WORKING

Steps	1000	500	200	100	50	30	20
IoU	64.41	63.32	64.28	65.71	61.32	59.43	*NW*
$F_a\downarrow$	8.58	8.13	7.96	5.80	2.59	5.61	*NW*
P_d	93.84	92.12	93.15	93.15	89.38	91.10	*NW*

The ROC curve of the IRSTD algorithm is special because of the large initial derivative. The response of true positive rate (TPR) to false positive rate (FPR) is very fast, but it tends to converge when FPR is about 5×10^{-5}. Therefore, in this case, AUC is almost only affected by TPR. ROC acts on the evaluation of pixel-level Fa is reflected in the horizontal axis, but the order of magnitude of the horizontal axis (10^{-5}) determines that it will not have a significant impact on AUC. Nevertheless, IRSTD-Diff represents strong competitiveness, second only to UIUNet. IRSTD-Diff has better performance than unlearnable methods and nonpertinent-designed generative methods. It also has competitiveness compared with the discriminant methods, such as ISNet and DNANet, which are recognized for their excellent performance.

C. Qualitative Comparisons

Qualitative results are obtained by different methods on NUAA-SIRST and IRSTD-1 k datasets. Most methods prefer to find relatively bright spots in the infrared images, owing to the dataset propensity. This brings numerous false alarm. Our IRSTD-Diff makes a better penalty for such a tendency, as shown in Figs. 8 and 9, thus achieving a significantly low false alarm. Meanwhile, its performance in missed detection and pixel-level edge fitting is competitive. UIUNet demonstrates a relatively coarse fit to the edges, presenting targets in blocklike pixel clusters. In contrast, DNANet and ISNet focus on target edges without blocklike clusters, but suffer from excessive smoothness, which impacts pixel-level performance. Our approach achieves a better compromise.

The proclivity of results from diffusion-based methods exhibits fundamental similarity, as shown in Fig. 10. These methods evince significantly fewer false alarms compared to the discriminative models in Fig. 9. Thereby, the effectiveness of our approach in addressing the hypothesis problem (Section I) is confirmed. However, there is a massive amount of miss detection, with lower Pd than state-of-the-art discriminative methods by more than 10%. Naturally, this results in substantial pixel-level deficiencies. The subpar pixel-level performance is not due to low accuracy in individual samples. On the contrary, the detected samples are precise in pixel-level. Given the prevalence of miss detection, resulting in 0 IoU for a considerable number of samples, the pixel-level performance is inevitably impacted. Our approach effectively averts this predicament, thereby facilitating the feasibility of the diffusion-based paradigm in IRSTD.

V. ABLATION STUDIES

We performed ablation studies on the IRSTD-1 k dataset, to verify the effectiveness of the latent space encoder I', low-frequency isolation in the wavelet domain and baseline, as shown in Table IX. All the above modules can improve the performance of the baseline.

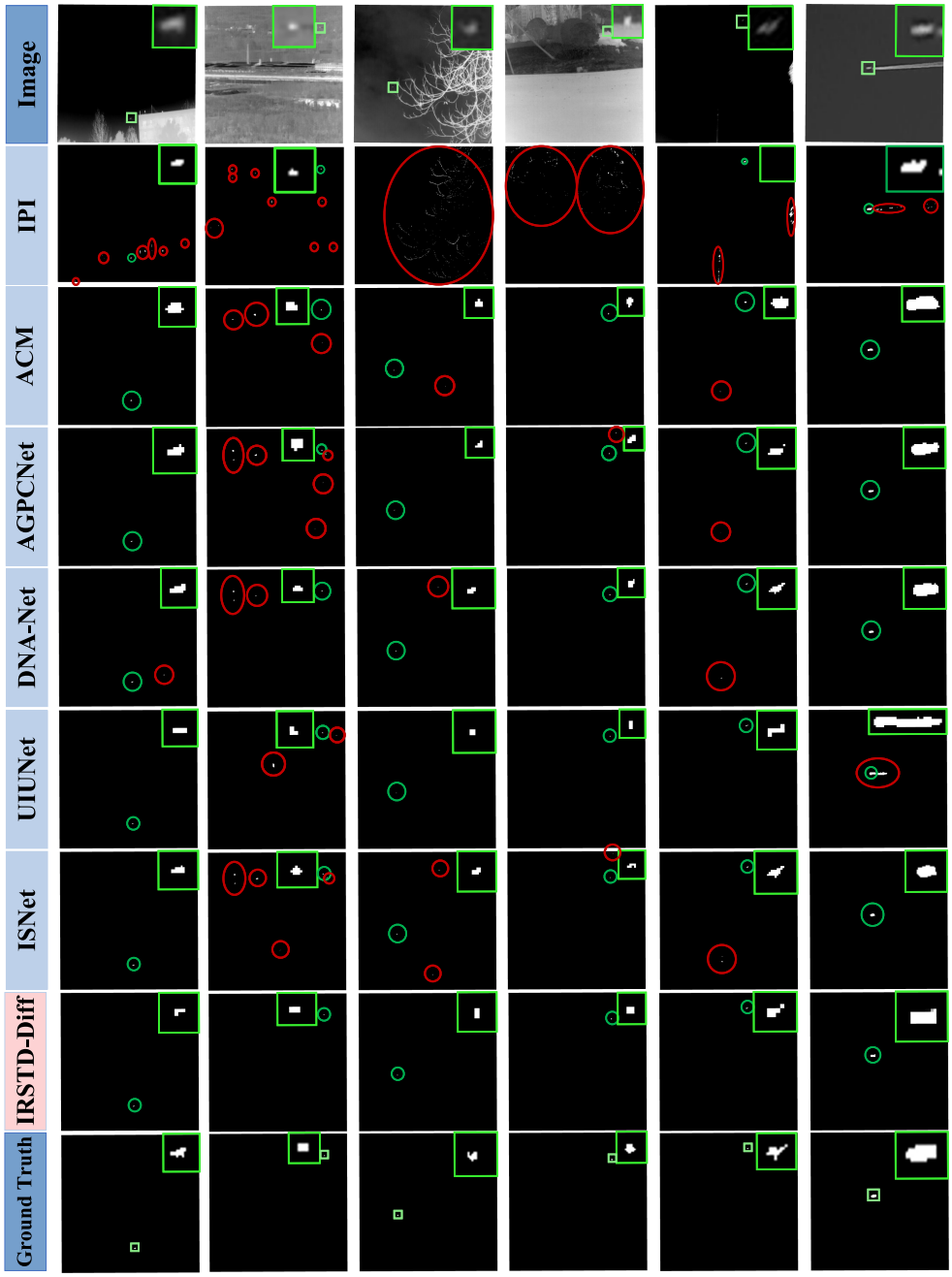

Fig. 8. Qualitative results obtained by different discriminative methods on NUAA-SIRST and IRSTD-1 k datasets. Enlarged targets are shown in the right-top corner. Circles in green, red, and yellow represent correctly detected targets, false alarm, and miss detected targets, respectively.

A. Effectiveness of the Latent Space Encoder $\mathcal{I}'$ and LIW Modules

If only the noise estimation network is leveraged for IRST mask generation, the pixel and target-level performance is inferior, as shown in Table IX # 2. After introducing the latent space embedding $\mathcal{I}'$ obtained by the CE, the IoU achieved an improvement of 5.42% (from 58.85% to 64.27%) due to more accurate low-level information, as shown in Table IX # 3. Moreover, the target-level performance has achieved a qualitative improvement, with the Fa reduced by 6.89×10^{-6} and the Pd increased by 4.46%. As shown in Table IX # 4, after the introduction of LIW, the pixel-level performance increased by 1.42%, and the Pd increased significantly (2.05%), which quantitatively demonstrated the effectiveness of the noise isolation operation.

We conducted ablation experiments to remove the noise estimation network, as shown in Table IX # 0 and # 1. Reasonable pixel-level and false alarm rate performance can still obtained, but the Pd is slightly insufficient. LIW module can improve the generative method, but it has a negative effect on the discriminative CE module. As is shown in Table IX # 0, #

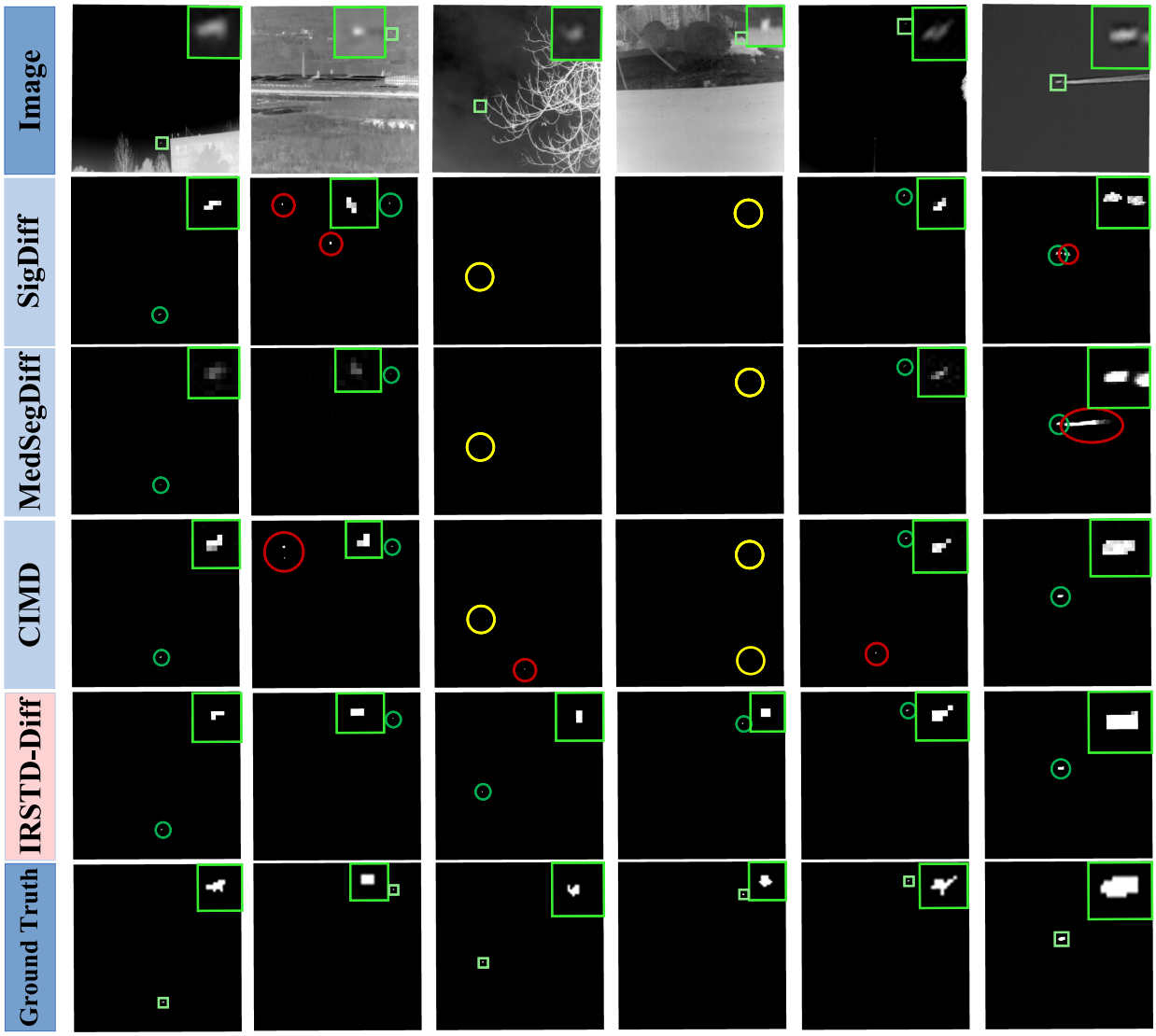

Fig. 9. Qualitative results obtained by diffusion-based generative methods on NUAA-SIRST and IRSTD-1 k datasets. Enlarged targets are shown in the right-top corner. Circles in green, red, and yellow represent correctly detected targets, false alarm, and miss detected targets, respectively.

TABLE VII
UNDERSAMPLING AND OVERSAMPLING DURING INFERENCE; THE TRAINING STEP IS SET TO 100, AND WE STOCHASTICALLY CHOOSE A STATIONARY POSTERIOR PROBABILITY (60 000 STEPS) FOR INFERENCE

Steps	500	300	100	80	60	40
$IoU(\%)$	64.93	64.76	64.98	64.61	64.42	31.66
$Fa(10^{-6})\downarrow$	8.74	8.24	7.42	7.06	7.00	51.92
$Pd(\%)$	93.15	93.15	93.15	93.15	93.15	92.81

TABLE VIII
EFFECTS OF THE ENSEMBLE

Ensembles	$IoU(\%)$	$P_d(\%)$	$F_a(10^{-6})\downarrow$
5	64.86	93.15	7.19
3	64.55	93.15	7.31
1	64.98	93.15	7.42

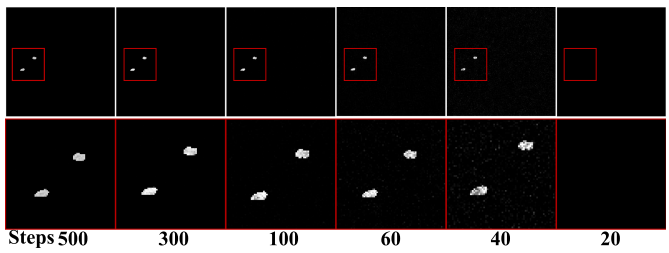

Fig. 10. Undersampling and oversampling with various steps. The second line is the magnification of the red box area in the first line.

2, and # 3, combined with the generative NE module can far surpass their individual existence, especially at the target-level, with the Fa reduced by 1.95 $\times$ 10^{-6} and the Pd increased by 2.40%. But IoU decreased slightly by 0.8%. In addition, the generative combinations have better target-level performance

TABLE IX
ABLATION: CONTRIBUTION OF OUR PROPOSED CE AND LIW

#	NE	CE	LIW	$IoU(\%)$	$F_a(10^{-6})\downarrow$	$P_d(\%)$
0		✓		65.07	7.63	88.70
1		✓	✓	60.30	8.74	90.07
2	✓			58.85	12.57	86.64
3	✓	✓		64.27	5.68	91.10
4	✓	✓	✓	65.69	5.80	93.15

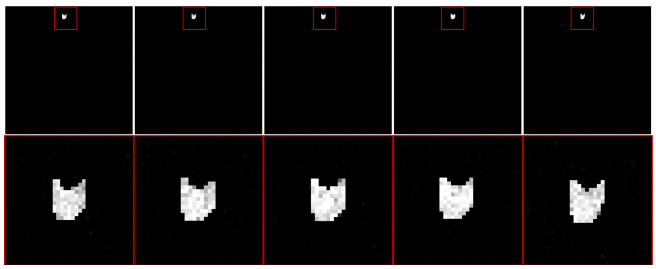

Fig. 11. Multiple samples under an identical posterior distribution. The second line is the magnification of the red box area in the first line. The results demonstrate the stability of sampling.

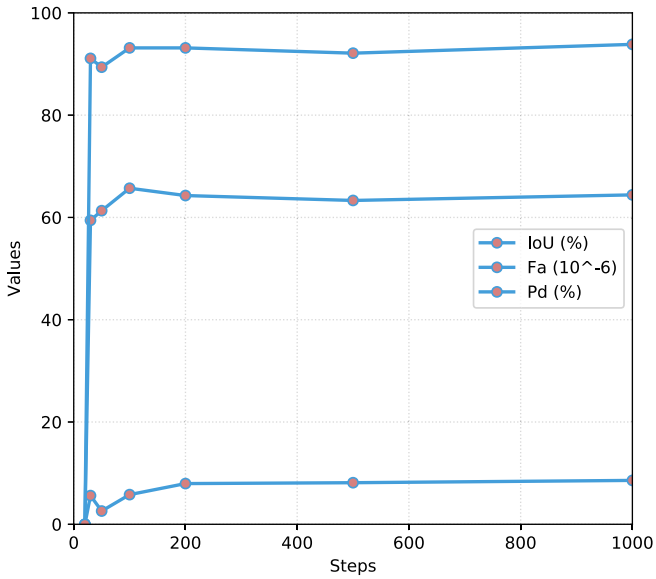

Fig. 12. Effects of the diffusion training steps. Small steps can achieve stupendous performance. When the step is greater than 100, continuing to increase the step is not beneficial to performance improvement, but oscillates in a small range.

than discriminative methods under the same framework, by comparing # 0, # 1 and # 3, # 4. This phenomenon demonstrates that the prior issue has been well solved.

The low-frequency isolation module in the wavelet domain (LIW), as shown in Fig. 6, conspicuously isolates the background radiation and obtains a cleaner infrared conditional encoder. It makes the conditional encoder network attach significance to the target area, which greatly reduces its attention dispersivity compared with no LIW module.

B. Impact of k-parameter in LIW

As shown in Table V, we discuss the value of k in the k-level 2-D *HDWT*. Wavelet transform, a double reduced resolution transform, makes the LIW enormously susceptible to k. An excessive wavelet transform with $k > 3$ will still lose several information because of the small target size, resulting in inferior performance. When $k = 1$, the module capacity is not enough to fit a better residual $\mathbf{R}$. In summary, the best k (resolution dependent) is 2 for 256 resolution.

C. Impact of Diffusion Training Steps

The mask data pattern is uncomplicated, thereupon the diffusion training steps can be smaller than the original configuration (1000) [28]. As shown in Table VI, taking 100 training steps has achieved a stupendous result. Under excessive steps, the evaluation metrics fluctuate, but there is no obvious improvement; and too few steps are not feasible.

Impact of Undersampling or Oversampling: The β is large at early steps, and the reverse process is intense, while the latter steps are converse [28]. This tendency is reflected in Fig. 10. To foreshorten the inference sampling time, undersampling can be adopted. As shown in Table VII and Fig. 12, 60 steps can achieve a reasonable performance. Oversampling does not bring any improvement.

D. Sample Stability and the Ensemble

The infrared small targets are small in size and simple in appearance, and consequently, the differential of multiple sampling results is not significant, as shown in Fig. 11. As quantitatively shown in Table VIII, the ensemble [59] is not necessary for the stability samples considering the sampling time.

VI. DISCUSSION

Previous deep learning-based IRSTD methods are discriminantive methods that have achieved reasonable pixel and target-level accuracy. However, these methods have serious target-level insensitivity, thus encounter a performance bottleneck in the false alarm rate and miss detection. In this study, we analyzed the principle of back propagation and model optimization comprehensively, and the problem is mitigated by the diffusion model-based generative approach.

IRSTD-Diff is an efficacious generative approach for IRSTD. Experimental results on two datasets demonstrate that IRSTD-Diff achieves superior performance against other diffusion-based and discriminative methods. However, diffusion-based methods have inferior inference sampling speed than discriminative methods, which entails additional sampling efficiency that will improve in the future. Besides, the diffusion-based methods are difficult to save the optimal results during training, which limits our approach.

Although the disadvantage of the DDPM is that it is not a real-time algorithm on medium/low-end equipment, IRSTD-Diff is friendly with acceptable parameters and edge deployment. Its computational complexity is $\frac{21.95 Steps}{Sampling Acceleration}$ GFLOPs.

At present, there are many methods to accelerate the diffusion model. For example, denoising diffusion implicit models (DDIMs) [60], PNDM [61], and EDM [62], etc., with non-Markovian ODE acceleration. DDIMs can produce high-quality samples $10\times$ to $50\times$ faster in terms of wall-clock time compared to DDPMs [60]. In addition, LCM [63], LCM-Lora [64], and other single-step knowledge distillation approaches that are compliant with the consistency constraint can also achieve it. At the inference of torch 2.0, if single-step inference LCM is leveraged, the average inference speed can be < 0.02 seconds per image.

Although many state-of-the-art performances have been achieved on challenging datasets, the pixel-level performances of the diffusion-based approaches are slightly inferior in some cases. More neural network modules and novel training paradigms designed for IRSTD should be proposed to solve the pixel-level-inferior in this generative framework. Moreover, is there any other generative paradigm that can be used to model posterior distribution more efficiently and pixel-level-superiorly? It is also feasible to explore another generative path, which is of great significance.

VII. Conclusion

In this article, we have presented a diffusion model framework for IRSTD, IRSTD-Diff, from the perspective of mask posterior distribution generating. The target-level insensitivity of the vanilla discriminative model is ameliorated. In addition, a low-frequency isolation module in the wavelet domain is designed, to reduce the impact of low-level interference infrared features on diffusion noise estimation. IRSTD-Diff is an efficacious generative approach for IRSTD. Experimental results on three datasets demonstrate that IRSTD-Diff achieves superior performance against other diffusion-based and discriminative methods. However, diffusion-based methods have inferior inference sampling speed than discriminative methods, which entails additional sampling efficiency improvement in the future. Besides, the diffusion-based methods are difficult to save the optimal results during training, which limits our approach.

References

[1] J. Lu, Y. He, H. Li, and F. Lu, "Detecting small target of ship at sea by infrared image," in *Proc. IEEE Int. Conf. Autom. Sci. Eng.*, 2006, pp. 165–169.

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

[3] P. Demosthenous, C. Pitris, and J. Georgiou, "Infrared fluorescence-based cancer screening capsule for the small intestine," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 2, pp. 467–476, Apr., 2016.

[4] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," in *Proc. Signal Data Process. Small Targets*, 1999, pp. 74–83.

[5] J.-F. Rivest and R. Fortin, "Detection of dim targets in digital infrared imagery by morphological image processing," *Opt. Eng.*, vol. 35, no. 7, pp. 1886–1893, 1996.

[6] J. Han, S. Moradi, I. Faramarzi, C. Liu, H. Zhang, and Q. Zhao, "A local contrast method for infrared small-target detection utilizing a tri-layer window," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1822–1826, Oct. 2020.

[7] J. Han et al., "Infrared small target detection based on the weighted strengthened local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1670–1674, Sep. 2021.

[8] 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.

[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] X. Wu, D. Hong, and J. Chanussot, "UIU-net: U-Net in U-net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2022.

[11] X. Yang, S. Li, S. Sun, and J. Yan, "Anti-occlusion infrared aerial target recognition with multisemantic graph skeleton model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[12] J. Lin, S. Li, L. Zhang, X. Yang, B. Yan, and Z. Meng, "IR-transdet: Infrared dim and small target detection with IR-transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023.

[13] 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.

[14] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2022.

[15] 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.*, 2022, pp. 877–886.

[16] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 516–520.

[17] H. Li, J. Yang, Y. Xv, and R. Wang, "ILNet: Low-level matters for salient infrared small target detection," 2023, *arXiv:2309.13646*.

[18] M.-Q. Le, T. V. Nguyen, T.-N. Le, T.-T. Do, M. N. Do, and M.-T. Tran, "MaskDiff: Modeling mask distribution with diffusion probabilistic model for few-shot instance segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 2874–2881.

[19] M. Vollmer, "Infrared thermal imaging," in *Computer Vision: A Reference Guide*. Berlin, Germany: Springer, 2021, pp. 666–670.

[20] H. Jian-Jiang, L. Zhao-Hui, and L. Wen, "Noise analysis of infrared image and muti-dim-small target's enhancement," *Infrared Technol.*, vol. 27, no. 2, pp. 135–138, 2005.

[21] M. Malfait and D. Roose, "Wavelet-based image denoising using a Markov random field a priori model," *IEEE Trans. image Process.*, vol. 6, no. 4, pp. 549–565, Apr. 1997.

[22] M. Jansen and A. Bultheel, "Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise," *IEEE Trans. Image Process.*, vol. 8, no. 7, pp. 947–953, Jul. 1999.

[23] V. Strela, P. N. Heller, G. Strang, P. Topiwala, and C. Heil, "The application of multiwavelet filterbanks to image processing," *IEEE Trans. Image Process.*, vol. 8, no. 4, pp. 548–563, Apr. 1999.

[24] N. Weyrich and G. T. Warhola, "Wavelet shrinkage and generalized cross validation for image denoising," *IEEE Trans. Image Process.*, vol. 7, no. 1, pp. 82–90, Jan. 1998.

[25] A. Haar, "Zur theorie der orthogonalen funktionensysteme," *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910.

[26] J. Li, P. Zhang, X. Wang, and S. Huang, "Infrared small-target detection algorithms: A survey," *J. Image Graph.*, vol. 25, no. 9, pp. 1739–1753, 2020.

[27] T. Zhang, S. Cao, T. Pu, and Z. Peng, "AGPCNet: Attention-guided pyramid context networks for infrared small target detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 4, pp. 4256–4261, Aug. 2023.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[29] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[30] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.

[31] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–14.

[32] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2649–2658.

[33] X. Liu et al., "More control for free! image synthesis with semantic diffusion guidance," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 289–299.

[34] A. Nichol et al., "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 16784–16804.
[35] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 36479–36494.
[36] Z. Feng et al., "ERNIE-ViLG 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10135–10145.
[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
[38] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proc IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11461–11471.
[39] M. Zhao, F. Bao, C. Li, and J. Zhu, "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 3609–3623.
[40] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.
[41] W. Xiang, H. Yang, D. Huang, and Y. Wang, "Denoising diffusion autoencoders are unified self-supervised learners," in *Proc. Int. Conf. Comput. Vision*, 2023, pp. 15765–15766.
[42] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," in *Int. Conf. on Learn. Representations*, 2022, pp. 1–15.
[43] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, "SegDiff: Image segmentation with diffusion probabilistic models," 2021, *arXiv:2112.00390*.
[44] J. Wu, H. Fang, Y. Zhang, Y. Yang, and Y. Xu, "MedSegDiff: Medical image segmentation with diffusion probabilistic model," in *Proc. Med. Imaging Deep Learn.*, 2024, pp. 1623–1639.
[45] J. Wu, R. Fu, H. Fang, Y. Zhang, and Y. Xu, "MedSegDiff-V2: Diffusion based medical image segmentation with transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 6030–6038.
[46] C. Ma et al., "DiffusionSeg: Adapting diffusion towards unsupervised object discovery," 2023, *arXiv:2303.09813*.
[47] J. Chen, J. Lu, X. Zhu, and L. Zhang, "Generative semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7111–7120.
[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.
[49] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.
[50] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested u-structure for salient object detection," *Pattern Recognit.*, vol. 106, 2020, Art. no. 107404.
[51] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[52] A. Vaswani et al., "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[54] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
[55] 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.
[56] 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.
[57] A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, and V. M. Patel, "Ambiguous medical image segmentation using diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11536–11546.
[58] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2018.
[59] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.
[60] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Representations*, 2021.
[61] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *Proc. Int. Conf. Learn. Representations*, 2022.
[62] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 26565–26577.
[63] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," 2023, *arXiv:2310.04378*.
[64] S. Luo et al., "LCM-LoRA: A universal stable-diffusion acceleration module," 2023, *arXiv:2311.05556*.

Haoqing Li received the B.Sc. degree in automation from Shandong University of Science and Technology, Qingdao, China, in 2021. He is currently working toward the M.Sc. degree in engineering with the Department of Control Science and Engineering, Beijing University of Technology, Beijing, China.

His research interests include computer vision, image semantic segmentation, deep learning, and their applications in infrared image processing.

Jinfu Yang received the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, in 2006.

During 2013–2014, he was a Visiting Scholar with the University of Waterloo, Canada. Since 2006, he has been with the Faculty of Information Technology and the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing, where he is currently a Professor of computer vision. His research interests include pattern recognition, computer vision, and robot navigation.

Dr. Yang is the Secretary General of the Beijing Association for Artificial Intelligence, and a Member of the Technical Committee of Computer Vision and Intelligent Robot of the Chinese Computer Federation (CCF CV and CCF TCIR).

Yifei Xu received the B.Sc. degree in measurement and control technology and instrument from Tianjin Polytechnic University, Tianjin, China, in 2021. He is currently working toward the M.Sc. degree in engineering with the Department of Control Science and Engineering, Beijing University of Technology, Beijing, China.

His research interests include deep neural network compression, computer vision, and robotic environmental awareness.

Runshi Wang received the B.Sc. degree in automation in 2021 from Beijing University of Technology, Beijing, China, where he is currently working toward the M.Sc. degree in engineering with the Department of Control Science and Engineering.

His current research interests include small target detection, image-to-image translation computer vision, and robot environment perception.