IRMamba: Pixel Difference Mamba with Layer Restoration for Infrared Small Target Detection

Mingjin Zhang, Xiaolong Li*, Fei Gao, Jie Guo*

Xidian University, China
{mjinzhang, fgao, jguo}@xidian.edu.cn, 23011210739@stu.xidian.edu.cn

Abstract

Infrared small target detection (IRSTD) focuses on identifying small targets in infrared images. Despite advancements with deep learning, challenges persist due to the IR long-range imaging mechanism, where targets are small, dim, and easily lost in noise and background clutter. Current deep learning methods struggle to suppress noise and background interference while preserving fine details, leading to missed detections and false alarms. To address these issues, we propose IRMamba, an encoder-decoder architecture featuring Pixel Difference Mamba (PDMamba) and a Layer Restoration Module (LRM). Specifically, PDMamba integrates the intensity and directional information of pixel differences between scanning positions and their central neighborhoods into the state equation of the state space model (SSM). This enhances target detail representation and suppresses background interference by capturing local 2D dependencies from a global perspective. In addition, LRM incorporates the double-depth image prior into the iterative convergence algorithm, and utilizes the inter-layer interrelationships to gradually reverse the separation of the target layer, achieving noise suppression and refined reconstruction of the image mask. Experiments conducted on multiple public datasets, including NUAA-SIRST, NUDT-SIRST, and IRSTD-1K, demonstrate the significant advantages of IRMamba over SOTA methods.

Introduction

Infrared small target detection (IRSTD) plays a crucial role in practical applications such as traffic management and maritime rescue (Cuccurullo et al. 2012; Law et al. 2016). Infrared sensors, which capture thermal radiation, provide valuable target information in difficult conditions such as fog, rain, or low-light environments. As a result, improving IRSTD performance has been a persistent focus in computer vision research (Zou et al. 2023). However, due to the remote imaging mechanism, targets in infrared images typically exhibit small and faint, while backgrounds are cluttered with significant noise, making targets easily obscured. Additionally, some background interferences in infrared images appear with high intensity and distinct shapes, creating strong contrast with surrounding pixels and increasing the likelihood of false alarms. These challenges underscore the ongoing technical difficulties in advancing IRSTD.

Current IRSTD methods fall into two categories: traditional algorithms and deep learning-based approaches. Early traditional methods, which dominated the field, treated IRSTD as a filtering and image enhancement problem (Han et al. 2020; Zhang et al. 2019, 2020). While effective in simple scenarios, these methods rely on prior knowledge and handcrafted features, limiting their accuracy when images deviate from their assumptions. In recent years, deep learning-based methods, particularly those using CNNs, have become widely adopted. CNNs expand the semantic receptive field through stacked convolutional layers to extract features, but they struggle to capture global context and low-level detail features—both crucial for small target detection. As a solution, hybrid methods combining Vision Transformers (ViTs) with CNNs (Zhang et al. 2022a; Yuan et al. 2024) have been proposed, leveraging ViTs' strength in modeling long-range dependencies and capturing global context.

However, these methods still have some design flaws: (1) ***Detail vs. Noise Balance.*** To address background noise and expand semantic range in infrared images, continuous downsampling is often used. But this process can lead to significant loss of target details due to their small and faint nature. To address this in IRSTD tasks, the U-Net architecture often employs skip connections to restore target detail during image reconstruction. Yet, balancing the need for rich target details with noise reduction remains challenging, often resulting in new noise accumulation in the decoder. (2) ***Complexity vs. Locality Balance.*** The self-attention mechanism in ViTs involves quadratic complexity, leading to low computational efficiency in these hybrid methods. While state space models (SSMs) like Mamba (Gu and Dao 2023) offer linear complexity and perform well in long sequence tasks, their method of flattening spatial data into 1D tokens disrupts local 2D dependencies of infrared small targets. This can cause the loss of critical local contrast features between targets and surrounding pixels in IRSTD tasks.

To tackle existing challenges, we propose an IRMamba network, applying the Mamba structure to IRSTD for the first time in this paper. Specifically, we design a Pixel Difference Mamba (PDMamba) module, which enhances local detail representation by integrating central pixel differentiation into the original state equation that captures in-

*Corresponding author

tensity information at scanning points. Guided by differences in central differential direction between background and small infrared targets, we devise a spatial distribution function within the state equation, allowing Mamba to selectively represent the current scanning position during global modeling. In addition, we develop a joint layer recovery method, which leverages inter-layer relationships to reverse the separation of target layers and reconstruct high-quality images from the stripped layer data. The Layer Restoration Module (LRM) employs distinct gradient descent paths for high-level semantic features and low-level detail features, enabling bidirectional information transfer. This approach balances noise reduction with detail preservation by cross-fusing gradient information from both perspectives, adjusting the gradient descent direction to prevent detail loss from over-filtering and noise introduction from over-supplementation. Experiments on public datasets, including NUAA-SIRST, NUDT-SIRST, and IRSTD-1K, validate the effectiveness of IRMamba and its significant advantages over representative state-of-the-art (SOTA) methods.

In summary, the contributions of this study are as follows:

- We introduce IRMamba, a novel design applied to the IRSTD field for the first time. In challenging benchmark tests, the proposed IRMamba outperforms the SOTA methods in both objective metrics and subjective evaluations, demonstrating its superior performance.
- We design PDMamba to capture neighbour information and target contour details by calculating the difference between scanned points and neighboring centers. Leveraging the spatial distribution of infrared images, we employ a weight function in the state equation to enhance target while suppressing background interference.
- We demonstrate the feasibility of separating clear layers from noisy image, followed by an image restoration strategy. The LRM alternates bidirectional gradient descent and cross-proximal mapping to suppress accumulated noise and recover the image mask.

Related Work

IRSTD Methods

Traditional IRSTD methods rely on image processing techniques or manual features, typically categorized into low-rank, human visual system-based, and filter-based methods. Low-rank methods, such as RIPT (Dai and Wu 2017) and NRAM (Zhang et al. 2018), adapt to low signal-to-clutter ratio (SCR) IR images but struggle with high false alarms in low-contrast regions. Human visual system-based methods, like TLLCM (Chen et al. 2013) and WSLCM (Han et al. 2020), are effective only in scenes with large objects and high background variability. Filter-based methods, including TOP-hat (Bai and Zhou 2010) and max-mean/max-median filtering (Deshpande et al. 1999), are limited to specific noise types and fail with complex backgrounds. In deep learning, IRSTD networks are divided into CNN-based and hybrid networks. CNN-based networks extract local features by stacking convolutional layers. For instance, MDvsFA (Wang, Zhou, and Wang 2019) utilizes GANs for a balance between target detection and false alarms, while ISNet (Zhang et al. 2022c), Dim2Clear (Zhang et al. 2023) and GCI-Net (Zhang et al. 2024b) focus on edge features and signal-to-noise ratio improvement, respectively. However, CNNs emphasize local features, while IRSTD also requires global context for effective false alarm suppression. Hybrid approaches combine CNNs with ViTs to integrate local and global information. For instance, IAANet (Wang et al. 2022) connects CNN patch outputs with transformers, and RKformer (Zhang et al. 2022a) employs transformers to highlight small IR targets and reduce background noise. Despite their effectiveness, they often suffer from high computational complexity due to ViTs' attention mechanism.

Vision Mamba

Recent advances in SSM, particularly Mamba(Gu and Dao 2023), have demonstrated the ability to model long-range dependencies while maintaining linear complexity, achieving strong performance in tasks like target detection and semantic segmentation (Zhang et al. 2024a). SSM, as a typical linear time-invariant system, transforms one-dimensional input signals $x(t) \in \mathbb{R}^L$ through intermediate hidden states $h(t) \in \mathbb{R}^N$, producing output $y(t) \in \mathbb{R}^L$. Mathematically, SSM is described by a linear ordinary differential equation:

$$h(t) = \overline{A} \cdot h(t-1) + \overline{B} \cdot x(t) \tag{1}$$

$$y(t) = Ch(t) + Dx(t) \tag{2}$$

These models incorporates various scanning strategies to enhance local perception. For example, Vim (Zhu et al. 2024) develops a selective horizontal scanning, while VMamba (Liu et al. 2024) adds vertical scanning to improve understanding of global context. LocalMamba (Huang et al. 2024) design a window-based local scanning strategy to capture dependencies efficiently while maintaining a global perspective. However, linear scanning curves often fail to capture dependencies between neighboring pixels in the same semantic region, limiting the model's ability to interpret spatial relationships. Given the small, dark targets and blurred edges typical in infrared imaging, IRSTD requires more precise pixel-level local visual information and edge detail.

Image Restoration

Image restoration aims to recover a high-quality image x from its degraded image Y. The degradation process is generally defined as:

$$Y = Ax + n, \tag{3}$$

$$\arg\min_x \parallel Y - Ax \parallel_2^2 + \lambda R(x), \tag{4}$$

where A represents the degradation matrix, n denotes additive noise, and λ is the hyperparameter of the weighted regularization term $R(x)$, aiding noise removal. Deep unfolding networks (DUNs) (Zhang, Van Gool, and Timofte 2020) are widely studied for solving image restoration problems. For example, DPDNN (Dong et al. 2019) uses denoising and back-projection modules for observation consistency, while DGUNet (Mou, Wang, and Zhang 2022) incorporates a gradient estimation strategy within a proximal gradient descent (PGD) algorithm for complex image restoration tasks. Given DUNs' powerful noise removal and restoration capabilities, we employ them in the joint task of layer restoration.

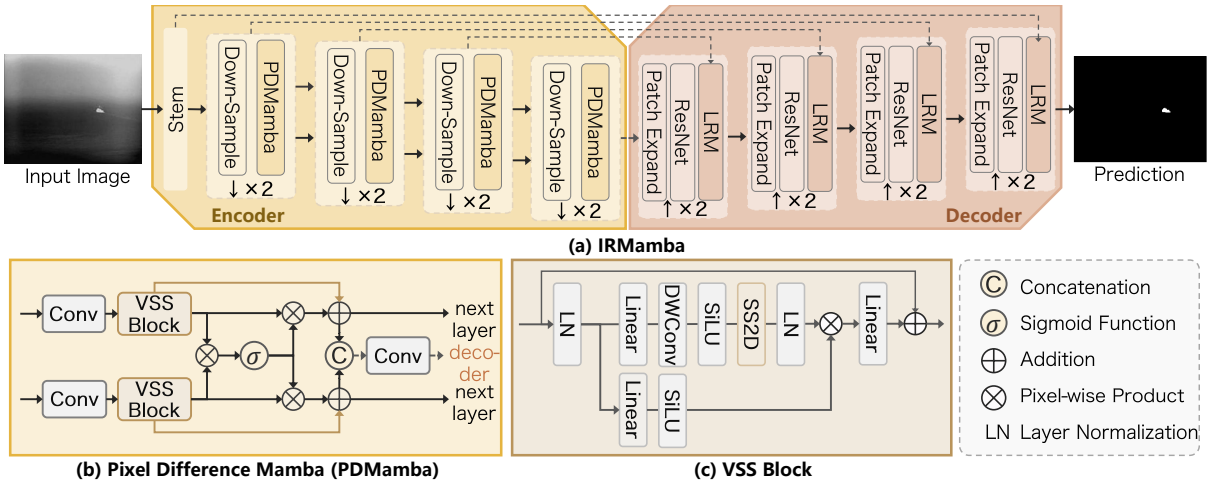

Figure 1: Overview of the proposed IRMamba encoder-decoder architecture.The encoder consists of two parallel stacked Convolutional and VSS Block modules, focusing on their respective tasks while sharing some mutual correlations through a gating mechanism for feature complementarity. The decoder sequentially stacks Patch Expanding, ResNet Blocks, and LRM to gradually recover the mask.

Methodology

Pixel Difference Mamba

In IRSTD tasks, the small size and low brightness of target objects require sharp local feature perception. While effective false alarm reduction demands leveraging global visual data to distinguish target from non-target regions in complex backgrounds. Mamba excels at modeling long-range dependencies and is widely used to extract global context and learn discriminative features in visual tasks. Nevertheless, the flattening of spatial tokens in Mamba hinders capturing tight pixel connections within the same semantic region. To enhance Mamba's local modeling capability, we design a PDMamba based on center pixel difference and a spatial distribution function.

Center Pixel Difference. In PDMamba's state equation, we enhance the state representation by incorporating pixel difference information with surrounding points, alongside the intensity of the current scanning point. This process is described by:

$$h(t) = \underbrace{\overline{A} \cdot h(t-1)}_{\text{History information}} + \underbrace{\overline{B} \cdot x(t)}_{\text{Intensity information}} + \underbrace{\sum_{i=1}^{n} \left[W_x(i) \cdot \Delta_x^{(o_i)} x(t) + W_y(i) \cdot \Delta_y^{(o_i)} x(t) \right]}_{\text{Neighbor information}} \tag{5}$$

$$\Delta_x^{(o_i^+)} = \frac{1}{2k+1} \sum_{k=0}^{i-1} (x(t+k) - x(t+k+1)), \tag{6}$$

$$\Delta_x^{(o_i^-)} = \frac{1}{2k+1} \sum_{k=0}^{i-1} (x(t-k+1) - x(t-k)). \tag{7}$$

Target detection accuracy is significantly improved by integrating intensity data from the current scanning position with the correlation information between its neighbors in the SSM state equation $h(t)$. Each pixel's brightness directly corresponds to the intensity of the scanning point, forming the basis for the $\overline{B} \cdot x(t)$ part in the SSM, which initializes or updates the state equation. Neighbor information is derived from the horizontal and vertical differential calculations $\Delta_x^{(o_i)} x(t)$ and $\Delta_y^{(o_i)} x(t)$ between the scanning position and central neighbor pixels, aiding in identifying mutual relationships and target contours. Here $\Delta_x^{(o_i)} x(t)$ and $\Delta_y^{(o_i)} x(t)$ are composed of $\Delta_x^{(o_i^+)}$, $\Delta_x^{(o_i^-)}$, $\Delta_y^{(o_i^+)}$, and $\Delta_y^{(o_i^-)}$, which represent the forward and backward smoothed differences in the horizontal and vertical directions, starting at the scanning point with a sliding window length of i. Adjusting n can expand the range of neighbourhood information acquisition, enhance the noise robustness, and adapt to different sizes of targets. Based on the small size of the target in infrared images, we set it to 2. Weighting coefficients $W_x(i)$ and $W_y(i)$ modulate the impact of differential information during state updates. This strategy enables PDMamba to maintain global perception while accurately capturing local dependencies, thereby improving detection accuracy.

Spatial Distribution Function. Background interference has a higher intensity and clearer shape, and its edges produce as significant a contrast with surrounding pixels as the target, making it challenging to distinguish targets from complex backgrounds. However, we found that the connection between the edges and the interior can cause the features to lose a certain degree of independence, leading to a deficiency in direction-specific gradient information. Combined with the above findings, we leverage Mamba's long-range modeling capability to better distinguish between background interference and the target, reducing false alarms.

Specifically, based on the above SSM state equation, we design a spatial distribution function using center difference direction information. When the scanning point aligns

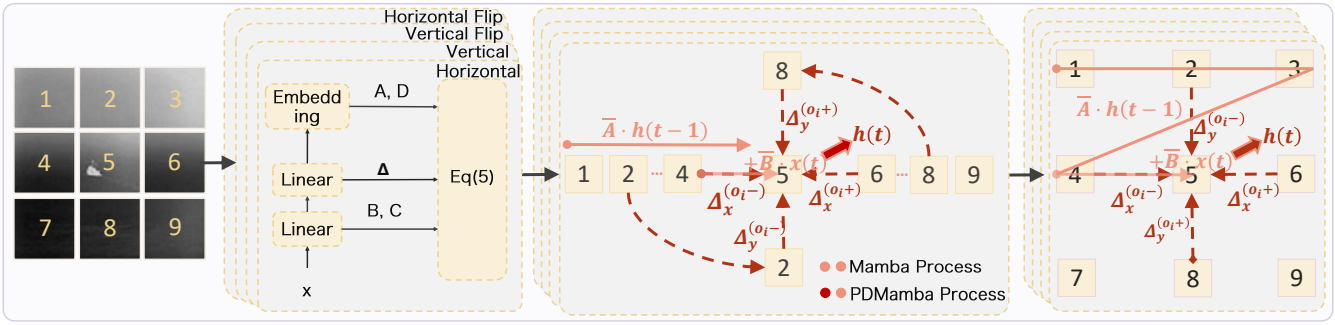

Figure 2: Illustration of scanning methods. We introduce difference operations into the equation of state, allowing Mamba to capture the difference information between the scanned points and the central neighbourhood during the scanning process. This approach enhances the ability to capture local dependencies.

with the spatial distribution of the infrared target, the weight value tends toward case 1 distribution; otherwise, it tends toward case 2 distribution. The state equation is given by:

$$h(t) = \overline{A} \cdot h(t-1) + \overline{B} \cdot x(t) + f\left(\Delta_x^{(o_n)}, \Delta_y^{(o_n)}; x(t)\right) \cdot \sum_{i=1}^{n} \left[W_x(i) \cdot \Delta_x^{(o_i)} x(t) + W_y(i) \cdot \Delta_y^{(o_i)} x(t)\right], \tag{8}$$

where the function f is defined as:

$$f\left(\Delta_x^{(o_n)}, \Delta_y^{(o_n)}; x(t)\right) = \begin{cases} \tanh\left(\frac{\Delta_x^{(o_n^+)} + \Delta_x^{(o_n^-)} + \Delta_y^{(o_n^+)} + \Delta_y^{(o_n^-)}}{4 \cdot \Theta(x(t), \sigma)} - 1\right) + \gamma \\ \gamma \end{cases},$$

where

$$\begin{cases} \Delta_x^{(o_n^+)}, \Delta_x^{(o_n^-)}, \Delta_y^{(o_n^+)}, \Delta_y^{(o_n^-)} > \Theta(x(t), \sigma) & \text{(for case 1)} \\ \text{otherwise} & \text{(for case 2)}, \end{cases} \tag{9}$$

$$\Theta(x(t), \sigma) = \sigma_n(t). \tag{10}$$

where the threshold Θ, is set by the local standard deviation $\sigma(t)$ of the current scanning point $x(t)$, indicating the average intensity variation in the current scanning point region. Additionally γ is set to 1 empirically.

In above, during the scanning process, if the shifted smoothing differences $\Delta_x^{(o_n^+)}$, $\Delta_x^{(o_n^-)}$, $\Delta_y^{(o_n^+)}$, and $\Delta_y^{(o_n^-)}$ centered at the current scanning point $x(t)$, with window length n, all exceed the local standard deviation $\sigma_n(t)$, it indicates significant contrast in any central direction, which is consistent with the spatial characteristics of small infrared targets. On the contrary, if this condition is not satisfied, the target does not comply with these characteristics. Accordingly, the PDMamba can enhance the representation of small infrared target features while leveraging Mamba's global modeling to suppress false alarms from background interference.

Layer Restoration Module

Due to the degradation in output quality caused by accumulated noise from the interaction of low-level and high-level features, we design a multi-objective optimization task aimed at addressing image denoising and image degradation restoration. Subsequently, we develop an LRM to iteratively solve this joint task, achieving refined image reconstruction and noise reduction through alternating Bidirectional Information Transmission Gradient Descent (BITGD) and Image Restoration Cross-Proximal Mapping (IRCPM) steps.

Layer Restoration Joint Mission. Within the framework of Deep Image Prior (DIP) (Ulyanov, Vedaldi, and Lempitsky 2017), tasks like image segmentation and denoising can be unified and interpreted as layer separation problems. Essentially, these tasks involve extracting and reconstructing different layers of information from complex input images. Based on this principle, Double DIP (Gandelsman, Shocher, and Irani 2019) proposes a reconstruction hypothesis: the separated image layers X and Y can be linearly combined to reconstruct the original image Z, indicating that complex images can be decomposed into different feature layers and then restored through appropriate weighted combinations:

$$Z = \beta \cdot Y + (1 - \beta) \cdot X, \tag{11}$$

where β is a weighting matrix.

Proof 1: According to this hypothesis, a clear image Y can be extracted from a noisy image Z by using the following weighted combination:

$$Y = \alpha \cdot X + (1 - \alpha) \cdot Z, \tag{12}$$

where $\alpha = 1/\beta$. Alternatively, Y can be derived by leveraging the mutual difference between the layers:

$$Y = \alpha \cdot X + (1 - \alpha) \cdot Z = \alpha \cdot (X - Z) + Z. \tag{13}$$

Proof 2: To derive the clear layer Y, we further integrate high-level semantic feature information (H) from the decoder and low-level detail feature information (L) from the skip connections. By utilizing the mutual relationships between these two layers, we can achieve the inverse separation of the clear layer Y:

$$Y = \alpha \cdot (L - H) + H. \tag{14}$$

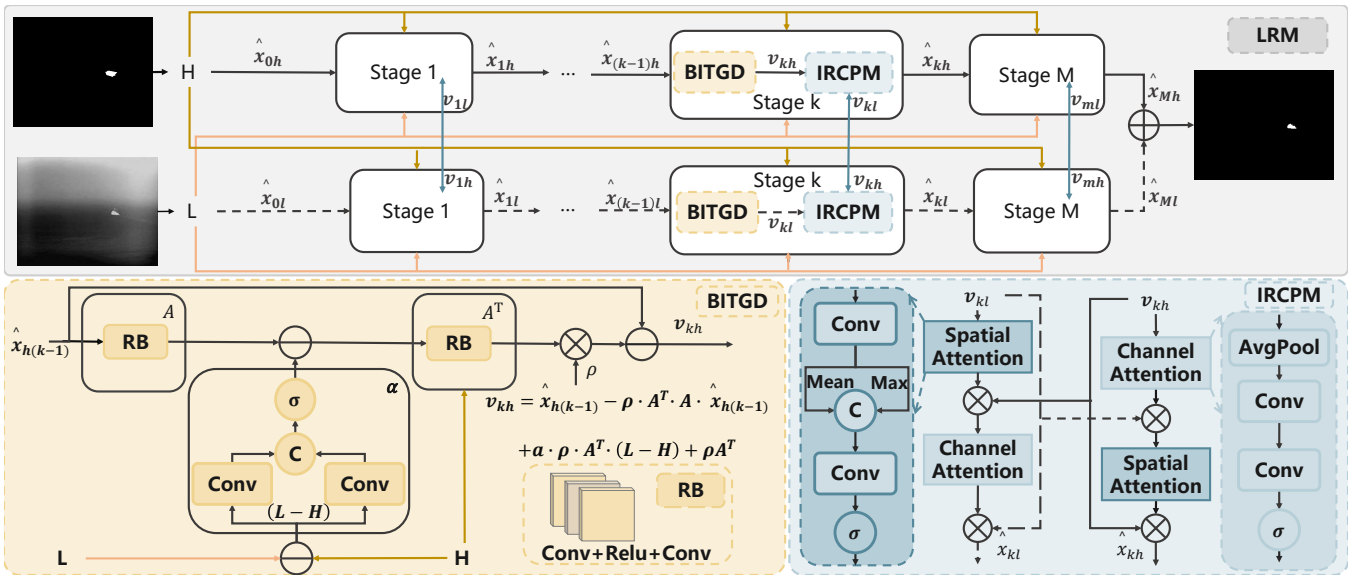

Figure 3: Overview of LRM. A multi-stage mask recovery strategy using Bidirectional Information Transmission Gradient Descen and Image Restoration Cross-Proximal Mapping in dual view.

Optimization: To refine the output quality of Y, we combine the image restoration theory formula (Eq. 3) and define the layer restoration process as a joint optimisation task.

$$Y = \alpha \cdot (L - H) + H = A \cdot x + n, \tag{15}$$

$$\arg\min_x \|\alpha \cdot (L - H) + H - Ax\|_2^2 + \lambda R(x). \tag{16}$$

We address this joint task by approximating the minimization problem as an iterative convergence problem. The minimization problem in the k-th iteration is defined as:

$$x_k = \arg\min_x \|x - (\hat{x}_{k-1} - \rho \nabla g(\hat{x}_{k-1}))\|^2 + \lambda R(x). \tag{17}$$

The iterative function is decomposed into two subproblems: gradient descent and proximal mapping. The LRM are comprised of multiple repeated stages, each containing a BITGD and an IRCPM, corresponding to the two subproblems in the algorithm's iterative steps. This innovative framework effectively solves the inverse problem of layer recovery and image restoration.

Bidirectional Information Transmission Gradient Descent. To iteratively solve the joint task and balance detail enhancement with noise removal, we propose a dual-view gradient descent strategy. It facilitates bidirectional gradient information transfer, enabling fine restoration and efficient noise removal within the same iteration.

Detail Restoration Perspective:

$$Y = \alpha_h \cdot (L - H) + H. \tag{18}$$

In this perspective, the model first focuses on global semantic information and iteratively refines α_h to supplement details from low-level features.

Denoising Perspectives:

$$Y = \alpha_l \cdot (H - L) + L. \tag{19}$$

In this perspective, the model first captures the edges and textures of the target image, using high-level semantic features to guide the iterative solution for α_l to filter out low-level noise.

Based on the above layer recovery strategy, in iteration k, we design two gradient descent paths for solving Y:

High-Level Path v_{kh}:

$$v_{kh} = \hat{x}_{h(k-1)} - \rho A^T A \hat{x}_{h(k-1)} + \rho A^T (\alpha_h \cdot (L - H)) + \rho A^T H, \tag{20}$$

Low-Level Path v_{kl}:

$$v_{kl} = \hat{x}_{l(k-1)} - \rho A^T A \hat{x}_{l(k-1)} + \rho A^T (\alpha_l \cdot (H - L)) + \rho A^T L, \tag{21}$$

where $\hat{x}_{k-1}$ is the previous stage's output, representing the current best estimate, ρ denotes the step size parameter controlling the update magnitude, A^T is the adjoint operator of A, used to map observation data back to the image domain.

Image Restoration Cross-Proximal Mapping. The layer restoration process utilizes two distinct strategies to achieve optimal results. Thus, IRCPM is introduced to merge key gradient information from low-level detail features, which enhance detail supplementation, and high-level semantic features, which filter low-level noise. It optimizes gradient descent by preventing excessive noise reduction that could lead to loss of detail, while also avoiding excessive detail enhancement that might introduce background noise.

$$\hat{x}_{kl} = v_{kl} \times (CA(SA(v_{kl}) \times v_{kh})), \tag{22}$$

$$\hat{x}_{kh} = v_{kh} \times (SA(CA(v_{kh}) \times v_{kl})), \tag{23}$$

where CA and SA stand for the channel and spatial attention, respectively. The update steps are repeated to incrementally approach the high-quality target layer. Leveraging the nonlinear mapping of deep networks and iterative optimization convergence, clear details are restored while noise is removed, ensuring high-quality layer restoration.

Method	Params ↓	Interface ↓	Type	NUDT-SIRST			IRSTD-1k			NUAA-SIRST		
				IoU ↑	P_d ↑	F_a ↓	IoU ↑	P_d ↑	F_a ↓	IoU ↑	P_d ↑	F_a ↓
IPI (Gao et al. 2013)	-	-	Traditional	30.93	81.98	17.99	27.92	81.37	16.18	1.09	87.05	30467
Top-Hat (Bai and Zhou 2010)	-	-	Traditional	22.40	89.90	174.1	10.06	75.11	1432	1.508	79.74	16456
MSLSTIPT (Sun, Yang, and An 2021)	-	-	Traditional	8.34	47.40	881	11.43	79.03	1524	1.080	0.052	8183
ACM (Dai et al. 2021a)	**0.52**	**0.010**	CNN-based	68.90	97.05	11.29	62.41	91.44	35.58	70.77	93.08	3.7
ALCNet (Dai et al. 2021b)	0.54	0.011	CNN-based	81.40	96.51	9.26	62.05	90.58	21.78	74.31	73.12	20.21
FC3-Net (Zhang et al. 2022b)	6.90	0.031	CNN-based	78.56	93.86	23.92	65.07	91.54	15.55	72.44	98.14	10.85
ISNet (Zhang et al. 2022c)	1.08	0.048	CNN-based	81.77	96.3	44.47	68.77	95.56	15.39	80.02	99.02	4.61
DNA-Net (Li et al. 2022)	4.7	0.023	CNN-based	88.19	98.62	9.00	69.38	93.3	11.66	79.26	98.48	2.3
UIU-Net (Wu, Hong, and Chanussot 2022)	50.54	0.032	CNN-based	92.19	97.77	15.44	69.96	93.98	22.07	78.25	97.45	2.296
Dim2Clear(Zhang et al. 2023)	5.71	0.024	CNN-based	81.37	96.23	9.17	66.34	93.75	20.93	77.29	99.10	6.72
GCI-Net (Zhang et al. 2024b)	0.71	0.021	CNN-based	92.43	98.25	8.96	67.75	93.89	12.84	78.81	99.34	2.108
RKformer (Zhang et al. 2022a)	29.00	0.065	Hybrid	92.25	96.86	6.58	64.12	93.27	18.65	77.24	99.11	1.58
IAANet (Wang et al. 2022)	14.05	0.092	Hybrid	90.22	97.26	8.32	66.25	93.15	14.20	74.22	93.53	22.70
SCTransNet (Yuan et al. 2024)	11.19	0.048	Hybrid	94.09	98.62	4.29	68.03	93.27	10.74	77.50	96.95	13.92
IRMamba (Ours)	10.51	0.044	Mamba	**95.18**	**99.26**	**1.309**	**70.04**	**95.81**	**5.92**	**80.2**	**99.36**	**0.887**

Table 1: Comparison with existing IRSTD approaches on the NUDT-SIRST, IRSTD-1k, and NUAA-SIRST datasets. The evaluation metrics are Params(M), Interface(s), IoU (10^{-2}), P_d (10^{-2}), and F_a (10^{-6}).

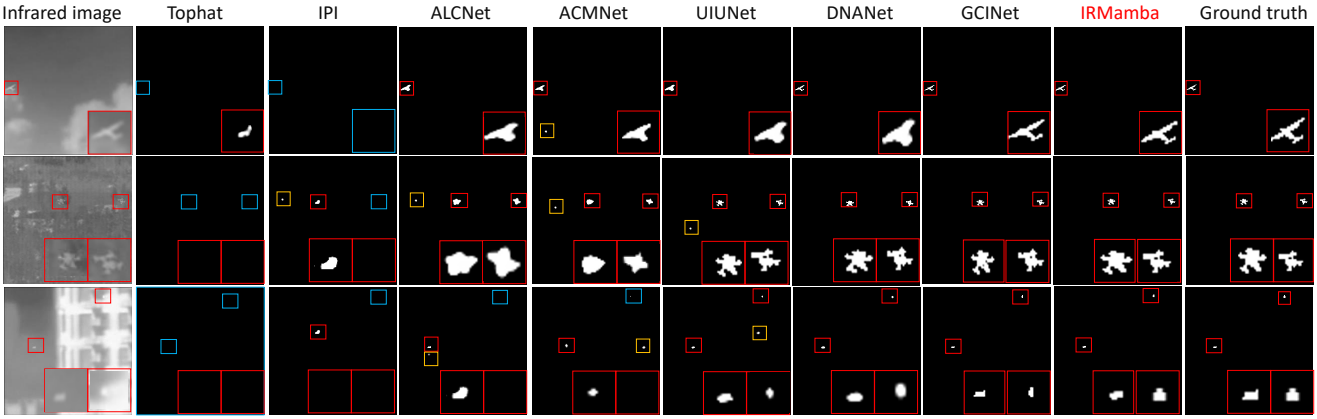

Figure 4: Visual results of different IRSTD methods. The boxes in red, yellow, and blue represent correct, missed, and false detections, respectively. Close-up views are shown in the corners.

Experiments

Experimental Details

Dataset: We use the NUAA-SIRST (Dai et al. 2021a), IRSTD-1k (Zhang et al. 2022c), and NUDT-SIRST (Zhang et al. 2022c) datasets for evaluation. These include 427 and 1,000 real infrared images with one or more small targets, and 1,327 synthetic infrared images of small targets in NUDT-SIRST. All images are resized to 256×256. For each dataset, we split the IR images into three disjoint subsets: 50% for training, 30% for validation, and 20% for testing.

Implementation Details: We utilize the Adam optimizer with an initial learning rate of 0.001 and employ the Cosine Decay Learning Rate Scheduler to train IRMamba model over 600 epochs with a batch size of 8 on a single Nvidia GeForce 4090 GPU with 24 GB of memory. For comparison, we include CNN-based methods (ISNet (Zhang et al. 2022c), GCINet (Zhang et al. 2024b), UIUNet (Wu, Hong, and Chanussot 2022), DNANet (Li et al. 2022), Dim2Clear (Zhang et al. 2023), ALCNet (Dai et al. 2021b), ACMNet (Dai et al. 2021a)), hybrid methods (RKformer (Zhang et al. 2022a), SCTransNet (Yuan et al. 2024), IAANet (Wang et al. 2022)), and traditional methods (Top-Hat (Bai and Zhou 2010), IPI (Gao et al. 2013)). We assess the performances using Intersection over Union (IoU), Probability of Detection (Pd) and False-Alarm Rate (Fa).

Main Results

Quantitative results: As shown in Tab. 1, traditional methods exhibit limitations in handling challenging cases. CNN-based methods struggle to adequately focus on target detail information, resulting in lower IoU scores. IRMamba outperforms Transformer-based models in terms of parameters, inference time, and performance. As resolution increases, the advantages of IRMamba become more apparent. At 1024x1024 resolution, its inference speed is more

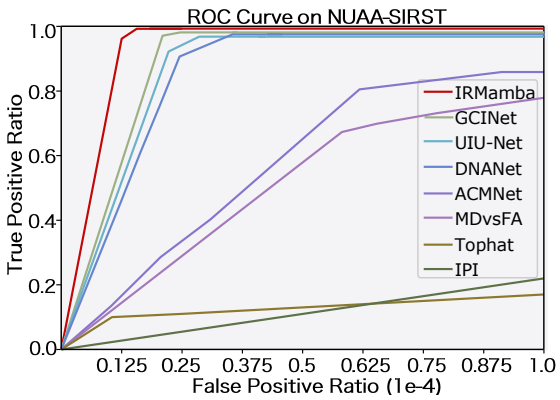

Figure 5: ROC curves of different methods.

Methods	IoU ↑	P_d ↑	F_a ↓
w/o PDMamba / LRM	91.09	94.75	9.44
w/o LRM	92.96	97.87	2.65
w/o PDMamba	92.28	97.01	1.98
IRMamba	**95.18**	**99.26**	**1.309**

Table 2: Ablation study of the proposed method.

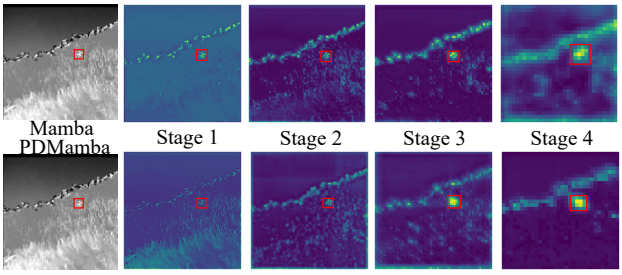

Figure 6: Visual comparison of PDMamba and Mamba.

than twice that of SCTransNet and RKformer, while IAANet faces memory overflow.

Visual Results: Fig. 4 presents detection results using IRMamba compared to other IRSTD methods. The superior performance of IRMamba is credited to PDMamba and LRM. These results highlight IRMamba's capability to effectively extract structural information of the target in challenging scenarios, such as background interference.

ROC: As shown in Fig. 5, our IRMamba significantly outperforms the other methods, where the Area Under the ROC Curve (AUC) of our IRMamba is larger than others.

Ablation Study

Impact of components: To evaluate IRMamba's components, we conducted ablation studies on the NUDT-SIRST dataset using a double encoder U-Net as a baseline. We systematically introduce the PDMamba and LRM modules. As illustrated in Tab. 2, removing PDMamba significantly reduces IoU and Pd scores, while increasing Fa scores. Additionally, Fig. 6 shows that with PDMamba as the encoder,

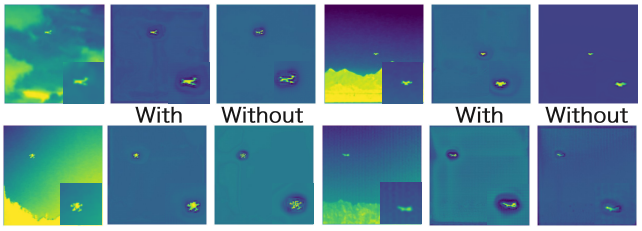

Figure 7: Visual comparison with and without LRM.

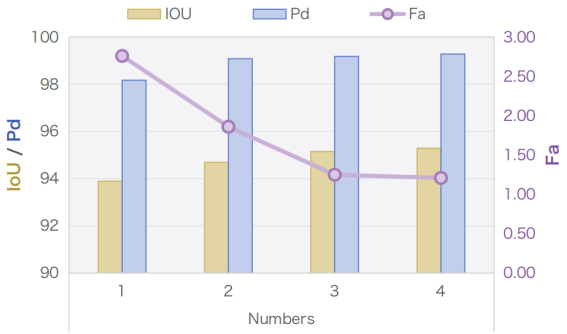

Figure 8: Ablation study on different numbers of LRM.

target features have higher, more concentrated intensity at each stage, while background interference features are significantly weaker. These results confirm PDMamba's effectiveness in capturing local dependencies and suppressing background interference. Similarly, removing LRM from the decoder decreases IoU and increases Fa. Meanwhile, Fig. 7 shows that LRM effectively suppresses the background noise and acquires finer IR small target features. These results demonstrating that LRM significantly enhance IR small target detection accuracy and noise suppression.

Number of stages: We analyze the performance gains across stages 1, 2, 3, and 4. As Fig. 8 shows, performance improves with each additional stage, highlighting the effectiveness of the iterative network design. As the decoder refines the mask, low-level edge details and noise become more pronounced, increasing the demand on LRM for refinement and denoising. Thus, we use 1, 2, 3, or 4 stages of LRM in the decoder to balance performance and complexity.

Conclusion

This paper introduces IRMamba for IRSTD, featuring LRM and PDMamba to enhance performance. PDMamba employs a difference operation to capture local dependencies and improve feature extraction, while its spatial distribution function reduces false alarms by mitigating background interference. LRM integrates layer separation with image degradation restoration, using a restoration network that alternates bi-directional gradient descent and cross-proximal mapping to accurately reconstruct targets and reduce noise. Experiments shows that IRMamba surpasses SOTA methods in objective metrics and subjective evaluations.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272363, Grant 92470108; in part by the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant 2021QNRC001; in part by the Joint Laboratory for Innovation in Satellite-Borne Computers and Electronics Technology Open Fund 2023 under Grant 2024KFKT001-1; in part by the Proof of Concept Foundation of Xidian University Hangzhou Institute of Technology under Grant No. GNYZ2023YL0301; in part by the Fundamental Research Funds for the Central Universities under Grant No. ZYTS24012.

References

Bai, X.; and Zhou, F. 2010. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition*, 43(6): 2145–2156.

Chen, C. P.; Li, H.; Wei, Y.; Xia, T.; and Tang, Y. Y. 2013. A local contrast method for small infrared target detection. *IEEE transactions on geoscience and remote sensing*, 52(1): 574–581.

Cuccurullo, G.; Giordano, L.; Albanese, D.; Cinquanta, L.; and Di Matteo, M. 2012. Infrared thermography assisted control for apples microwave drying. *Journal of food engineering*, 112(4): 319–325.

Dai, Y.; and Wu, Y. 2017. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE journal of selected topics in applied earth observations and remote sensing*, 10(8): 3752–3767.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021a. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 950–959.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021b. Attentional local contrast networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11): 9813–9824.

Deshpande, S. D.; Er, M. H.; Venkateswarlu, R.; and Chan, P. 1999. Max-mean and max-median filters for detection of small targets. In *Signal and Data Processing of Small Targets 1999*, volume 3809, 74–83. SPIE.

Dong, W.; Wang, P.; Yin, W.; Shi, G.; Wu, F.; and Lu, X. 2019. Denoising Prior Driven Deep Neural Network for Image Restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10): 2305–2318.

Gandelsman, Y.; Shocher, A.; and Irani, M. 2019. "Double-DIP": Unsupervised Image Decomposition via Coupled Deep-Image-Priors. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11018–11027.

Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; and Hauptmann, A. G. 2013. Infrared patch-image model for small target detection in a single image. *IEEE transactions on image processing*, 22(12): 4996–5009.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *ArXiv*, abs/2312.00752.

Han, J.; Moradi, S.; Faramarzi, I.; Zhang, H.; Zhao, Q.; Zhang, X.; and Li, N. 2020. Infrared small target detection based on the weighted strengthened local contrast measure. *IEEE Geoscience and Remote Sensing Letters*, 18(9): 1670–1674.

Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024. LocalMamba: Visual State Space Model with Windowed Selective Scan. *ArXiv*, abs/2403.09338.

Law, W.-C.; Xu, Z.; Yong, K.-T.; Liu, X.; Swihart, M. T.; Seshadri, M.; and Prasad, P. N. 2016. Manganese-doped near-infrared emitting nanocrystals for in vivo biomedical imaging. *Optics express*, 24(16): 17553–17561.

Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; and Guo, Y. 2022. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. VMamba: Visual State Space Model. *ArXiv*, abs/2401.10166.

Mou, C.; Wang, Q.; and Zhang, J. 2022. Deep Generalized Unfolding Networks for Image Restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17378–17389.

Sun, Y.; Yang, J.; and An, W. 2021. Infrared Dim and Small Target Detection via Multiple Subspace Learning and Spatial-Temporal Patch-Tensor Model. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5): 3737–3752.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. S. 2017. Deep Image Prior. *International Journal of Computer Vision*, 128: 1867 – 1888.

Wang, H.; Zhou, L.; and Wang, L. 2019. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8509–8518.

Wang, K.; Du, S.; Liu, C.; and Cao, Z. 2022. Interior Attention-Aware Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Wu, X.; Hong, D.; and Chanussot, J. 2022. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing*, 32: 364–376.

Yuan, S.; Qin, H.; Yan, X.; Akhtar, N.; and Mian, A. 2024. SCTransNet: Spatial-Channel Cross Transformer Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.

Zhang, H.; Zhu, Y.; Wang, D.; Zhang, L.; Chen, T.; and Ye, Z. 2024a. A Survey on Visual Mamba. *ArXiv*, abs/2404.15956.

Zhang, K.; Van Gool, L.; and Timofte, R. 2020. Deep Unfolding Network for Image Super-Resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3214–3223.

Zhang, L.; Peng, L.; Zhang, T.; Cao, S.; and Peng, Z. 2018. Infrared small target detection via non-convex rank approximation minimization joint l 2, 1 norm. *Remote Sensing*, 10(11): 1821.

Zhang, M.; Bai, H.; Zhang, J.; Zhang, R.; Wang, C.; Guo, J.; and Gao, X. 2022a. RKformer: Runge-Kutta Transformer with Random-Connection Attention for Infrared Small Target Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1730–1738.

Zhang, M.; Wang, N.; Li, Y.; and Gao, X. 2019. Deep Latent Low-Rank Representation for Face Sketch Synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10): 3109–3123.

Zhang, M.; Wang, N.; Li, Y.; and Gao, X. 2020. Neural Probabilistic Graphical Model for Face Sketch Synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7): 2623–2637.

Zhang, M.; Yue, K.; Li, B.; Guo, J.; Li, Y.; and Gao, X. 2024b. Single-Frame Infrared Small Target Detection via Gaussian Curvature Inspired Network. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13.

Zhang, M.; Yue, K.; Zhang, J.; Li, Y.; and Gao, X. 2022b. Exploring Feature Compensation and Cross-level Correlation for Infrared Small Target Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1857–1865.

Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; and Guo, J. 2022c. ISNET: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 877–886.

Zhang, M.; Zhang, R.; Zhang, J.; Guo, J.; Li, Y.; and Gao, X. 2023. Dim2Clear Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *ArXiv*, abs/2401.09417.

Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; and Ye, J. 2023. Object detection in 20 years: A survey. *Proceedings of the IEEE*.