

## Article

# HMCNet: A Hybrid Mamba–CNN UNet for Infrared Small Target Detection

Bolin Li <sup>1,2,3</sup> , Peng Rao <sup>1,2</sup>, Yueqi Su <sup>1,2</sup>  and Xin Chen <sup>1,2,\*</sup>

<sup>1</sup> Key Laboratory of Intelligent Infrared Perception, Chinese Academy of Sciences, Shanghai 200083, China; libolin22@mailsucas.ac.cn (B.L.); peng\_rao@mail.sitp.ac.cn (P.R.); suyueqi@mail.sitp.ac.cn (Y.S.)

<sup>2</sup> Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China

<sup>3</sup> University of Chinese Academy of Sciences, Beijing 100049, China

\* Correspondence: chenxin@mail.sitp.ac.cn

**Abstract:** Using infrared technology to accurately detect small weak targets is crucial in various fields, such as reconnaissance and security. However, the infrared detection of small weak targets is challenged by complex backgrounds, tiny target sizes, and low signal-to-noise ratios, which significantly increase the difficulty of detection. Early studies in this domain typically utilized manually designed feature-extraction methods that performed inadequately in the presence of complex backgrounds. While advancements in deep learning have spurred rapid progress in this field, with CNN models effectively enhancing the detection performance, the problem of small weak target features being lost persists. HMCNet, which employs a hybrid architecture combining a state space model and a CNN, is proposed in this paper; its hybrid architecture demonstrates the capacity to extract the local features and model the global context, facilitating superior suppression of complex backgrounds and detection of small weak targets. Our experimental results on the publicIRSTD-1k dataset and our own MISTD dataset indicate that, compared to the current mainstream methods, the method proposed achieves better detection accuracy while maintaining high-speed inference capabilities, thus validating the rationality and effectiveness of this research.

**Keywords:** infrared small target; segmentation; attention mechanism; state space model



Academic Editor: John Trinder

Received: 21 November 2024

Revised: 19 January 2025

Accepted: 26 January 2025

Published: 29 January 2025

**Citation:** Li, B.; Rao, P.; Su, Y.; Chen, X. HMCNet: A Hybrid Mamba–CNN UNet for Infrared Small Target Detection. *Remote Sens.* **2025**, *17*, 452. <https://doi.org/10.3390/rs17030452>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Due to the ability of infrared sensors to capture infrared radiation emitted or reflected by objects and convert it into visible images, infrared small target detection is widely applied across military reconnaissance, security surveillance, forest fire prevention, and assistance systems for night driving [1], contexts in which target identification in low-light or dark environments is required [2]. Infrared small target detection has advantages over radar and visible light detection, such as good concealment, strong penetration, and a passive detection mechanism. However, despite these many advantages, infrared small target detection also faces some challenges in practical applications. Firstly, due to the long detection distances and actual sizes of the small targets involved, these small targets are represented by very few pixels in the images, complicating effective feature extraction. Additionally, the signal-to-noise ratio (SNR) is often very low because the signals from small targets are typically weak and easily masked by background noise. These difficulties affect the accuracy of detection. There are high requirements for real-time performance in infrared small target detection scenarios, especially in applications such as vehicle night vision systems, where potential hazards need to be identified quickly and accurately in dynamic

scenes, leading to high demands in terms of the processing speed of the algorithms. On these bases, further developments in infrared small target detection (ISDT) are pivotal [3,4].

The traditional methods for detecting small targets in infrared imagery have followed a model-driven approach, in which mathematical models for problem-solving are established based on theory. These include several representative methods. Filter-based methods [5–9] enhance the target features and suppress the background through the filter design, thereby improving the signal-to-noise ratio. Methods that utilize local contrast mechanisms [10–14] increase the intensity of weak signals to enhance the contrast, thus highlighting the targets. Methods involving the decomposition of sparse and low-rank tensors

The application of deep learning to the field of infrared detection of weak and small targets has become a popular direction within the research. Unlike the traditional model-driven methods, neural networks are data-driven and offer many advantages: end-to-end learning without the need for manually designed features, strong generalization, hardware acceleration, parallel processing capabilities, high precision, and robustness. Given their suitability for feature encoding and decoding, CNN networks excel at extracting hierarchical feature information. Their weight-sharing design makes them particularly good at handling translation invariance and recognizing local features. Within this field, the concept of target segmentation using a UNet has inspired many representative works. For example, Wang et al. [15] utilized the concept of adversarial learning in Generative Adversarial Networks (GANs) to achieve a balance between the two subtasks of missed detections (MDs) and false alarms (FAs). With each network focused on one subtask, task flexibility is improved. Weak and small target detection is also aided using semantic segmentation techniques, such as using a symmetrical encoder–decoder structure. Dai et al. published several important works in this vein. ACMnet [16] employs asymmetric contextual modulation, integrating low-level texture information and high-level semantic information to alleviate the issue of low-level feature loss. ALCnet [17] uses the method of cyclically shifting the feature maps, thus surpassing the limitations of the receptive field of the convolutional kernels to facilitate long-distance contextual interactions. Li et al. proposed DNANet [18], which enhances the robustness of the network through the dense, nested UNet architecture to promote interaction and fusion of the features between layers. Wu et al. introduced UIUnet [19], in which the feature integration in the original UNet was improved by replacing its feature-extraction blocks with residual U-blocks. In addition, to ensure sufficient training data are available and promote the development of the field, certain research teams have provided public and open-source infrared target detection datasets [15,16,18,20,21].

Nevertheless, where CNNs are applied within this domain, the challenge of lower-level feature loss arises, leading to poorer performance in deeper networks and a tendency towards overfitting. Transformers, on the other hand, are increasingly popular neural network architectures that have also been successfully applied to image processing despite being conceived for natural language processing purposes initially. For example, Vision Transformers (ViTs) have been used for image recognition [22], the Swin Transformer serves as a universal backbone network in various visual tasks [23], and they have also found extensive applications for processing remote sensing images [24]. Unlike CNNs, transformers treat images as a sequence of patches instead of naturally considering their spatial hierarchical structures. They are better at capturing global information, and the introduction of global dependencies somewhat mitigates the issue of lower-level information being diluted by higher-level information, thus addressing the aforementioned shortcomings of CNNs. Many studies have explored the use of hybrid transformer–CNN network architectures in image segmentation tasks [25–28]. IAAAnet, proposed by Wang et al., employs a transformer encoder to model the attention relationships within the target

regions, utilizing a range from coarse to fine [29]. Despite their improved capacity to handle long-range dependencies, transformers typically incur high computational costs; as the self-attention mechanism grows quadratically with the size of the input, high-resolution image processing requires a sharp increase in computational complexity.

Recently, state space models (SSMs), particularly the Structured State Space Column Model (S4), have served as an effective backbone for constructing deep networks, achieving optimal performance in analyzing continuous, long sequences [30]. Mamba further improves upon S4 by introducing a selection mechanism that allows the model to choose information that is relevant to the input. Where it has been coupled with a hardware-aware implementation, Mamba has surpassed transformers in tackling dense modalities such as language and genomics. Moreover, state space models have shown promising results in visual tasks, including image and video classification [31,32]. In the field of infrared small target detection, Chen et al. proposed MIM [33], which uses a nested Mamba model with inner and outer layers to extract the global information ingeniously. Given that image patches and image features can also be considered sequences [22,23], we have been inspired to explore the potential to use the Mamba model to enhance the capability of CNNs to model long distances.

Thus, we designed HMCNet, a UNet that integrates a hybrid Mamba–CNN module. This network adopts the symmetric encoder–decoder design within the UNet, where feature layers of the same depth in the encoder and the decoder are connected through skip connections to facilitate context interactions. For each feature extraction block, we designed a Hybrid Vision State Space Module (HY-SSM) to integrate the convolutional branch and the Mamba branch. The convolutional branch uses multi-scale hybrid modules and focus-aware modules to increase the receptive field and enhance the local feature extraction, while the Mamba branch imbues the model with efficient global attention capabilities. Thus, the model’s design not only retains the power of CNNs in feature extraction from local receptive fields but also introduces more computationally friendly global attention modules. Our experiments on both public and private datasets show this method’s effectiveness in achieving the best results and a faster inference speed.

The main contributions of this work can be summarized as follows:

- We propose HMCNet: a UNet that combines a CNN and the Mamba backbone network for small target segmentation using infrared, unifying the characteristics of CNNs in enhancing the local receptive fields and the advantages of state space models in terms of their low computational complexity and linear scalability.
- We designed a multi-scale fusion block (MSFB) and a focus-aware block (FAB). These modules enhance the extraction of the local features while increasing the local receptive field.
- The experimental results on the public IRSTD-1k dataset and the MISTD dataset we built indicate that the method proposed achieves superior detection accuracy and retains high-speed inference capabilities, thus validating the rationality and effectiveness of this research.

## 2. Related Work

### 2.1. Infrared Small Target Detection

Detecting weak and small targets using infrared has always been a key research focus. The traditional methods treat this problem as an anomaly detection problem using prior assumptions, modeling infrared images based on mathematical theory. Filter-based methods [5,6] extract the features by removing noise through convolution. Morphological methods [7] use binary operations to process the images, but they are easily affected by noise. Algorithms based on the human visual system (HVS) [10,34] enhance the local

contrast to obtain saliency maps using neighborhood differences, which enhance the signals and suppress noise but require the segmentation thresholds to be manually adjusted. Sparse matrix optimization methods [35–39] solve detection problems by turning them into optimization problems. Although these traditional methods are conceptually simple, they require manual parameter tuning, rely on prior knowledge, and have low adaptability.

In recent years, developments in deep learning and computer vision have also driven advancements in the field of infrared small target detection. Due to the small proportions of the targets within the entire image space (as defined by SPIE [40]), the detection bounding boxes often include background noise, leading to insufficient precision. Image segmentation methods provide greater accuracy than bounding box detecting in offering pixel-level classification information. The UNet [41–43] has achieved great success in the field of image segmentation, becoming one of the classic models in this domain. Many research teams have applied image segmentation models within the field of infrared small target segmentation and achieved good results. Dai et al. [16,17] enhanced the interaction between low-level and high-level information in UNets and its retention by introducing asymmetric contextual modulation. ISNet [20] introduced prior information on contours into the UNet, achieving cross-layer feature fusion, which aided in reconstructing the targets. AMFUNet [44] utilized the full-scale connection features of U-Net3+ [43] to fully integrate the feature information from each layer. In its incorporation of multiple cascading small UNets with carefully designed feature fusion mechanisms, UTUNet [19] exhibited enhanced feature-extraction capabilities. Beyond networks that incorporate the UNet structure, there has been a host of other outstanding work. For instance, Kim et al. [45] used Generative Adversarial Networks to simulate infrared images and targets, therefore improving the accuracy of target detection. Through its integration of local contrast mechanisms and multi-scale receptive fields in a lightweight network, RDIAN [21] achieved excellent results. However, although CNN-based networks can adapt to varying scenarios and offer convenient features, such as lacking the need for manual design, finding the right balance between the quantity of parameters and accuracy is often challenging, and the issue of low-level details vanishing also limits further breakthroughs in this domain.

## 2.2. Attention Mechanisms

Since its initial application in the field of natural language processing, self-attention [46] has attracted a lot of research activity. It later gained popularity in computer-vision-related fields such as detection and segmentation. Squeeze and Excitation Networks (SENet) [47] were the first CNN models to widely adopt attention mechanisms, namely focusing on channel dimension attention, with a global pooling module used to perform a weighted selection of the channel feature maps. In paying attention to regions of interest in both the spatial and temporal dimensions, the Convolutional Block Attention Module (CBAM) [48] extended the SENet to the spatial dimension. Attention mechanism modules were initially designed for segmenting large target objects in images. Equally, due to the lack of texture and color information in infrared, small target detection in infrared differs from detection in traditional scenes. Therefore, applying attention mechanisms to infrared small target detection requires the use of carefully designed attention modules and feature-extraction networks. Indeed, some existing works have integrated attention mechanism modules into CNN feature-extraction modules for segmentation purposes [19].

In addition to attention in the CNN domain, the transformer model, which has gained immense popularity in recent years, is a prime example of a global attention mechanism. Notable examples of transformers for image tasks include Vision Transformers [22], and Swin Transformers [23], both of which have achieved remarkable results. Global attention can capture the relationships between different parts of an image, overcoming the

limitations of a CNN's local receptive fields. In CNNs, parameters such as the kernel size, depth, and stride need to be manually designed and adjusted, whereas transformers can automatically learn the attention weights for each position from the training data, partially reducing the complexity of the manual design. In deep CNN networks, information on small targets may be overwhelmed, while using global receptive fields can help mitigate the dilution of low-level information. However, transformers require a large amount of training data, and the complexity of global attention grows quadratically, making them less efficient than CNNs with shared weights in this regard. To leverage the advantages of both structures, many studies have explored hybrid architectures that combine transformers and CNNs [25,27,28], including recent work on their application to infrared small target segmentation. For example, IAA-net [29] uses a region proposal network to perform a rough background screening, while another branch models the attention for the target area, with both branches working together to complete the task. Liu et al. [49] applied transformer modules in both the encoder and the decoder to model the attention for the features encoded by the UNet. In hybrid models, reducing the computational burden of the transformer and improving the parallelism are among the key directions for their optimization.

### 2.3. State Space Models

State space models (SSMs) have gradually become popular in deep learning over the past two years, especially the Structured State Space Model (S4), which has demonstrated unprecedented superiority in handling long-sequence data. The core advantage of these models lies in their ability to efficiently capture long-range dependencies through state space equations while maintaining the linear computational complexity. This significantly enhances their performance in fields such as natural language processing and time series analysis. The Mamba model improves upon S4 further with the introduction of a dynamic selection mechanism that allows the model to flexibly choose the most relevant information based on the input. This is similar to an attention mechanism but is more efficient for dealing with long sequences than traditional attention models. Mamba also incorporates hardware-aware optimization strategies, allowing it to outperform the mainstream transformer models in complex tasks such as language modeling and genomics analyses. Moreover, the application of state space models has expanded into the field of computer vision [31,32]. In viewing the image patches or features as ordered data sequences in image classification and video analysis tasks, state space models can leverage this to model long-range dependencies, thereby significantly improving the performance of convolutional neural networks (CNNs) in these tasks. Future research directions may focus on integrating state space models with CNN architectures more efficiently to enhance the synergistic effects of their advantages for visual tasks [50–52].

## 3. Methods

### 3.1. Preliminaries

#### 3.1.1. The State Space Model

State space models (SSMs) [53] are utilized to depict the representation of the state of a sequence at various time steps and predict its subsequent state based on the input. They are commonly employed for linear time-invariant systems. A one-dimensional input signal  $x(t) \in \mathbb{R}$  is transformed into the predicted output signal  $y(t) \in \mathbb{R}$  through a hidden state  $h(t) \in \mathbb{R}^N$ . A differential equation representing the process above is as follows:

$$\begin{cases} h'(t) = Ah(t) + Bx(t) \\ y(t) = Ch(t) + Dx(t) \end{cases} \quad (1)$$

where  $A \in \mathbb{R}^{N \times N}$  is the state matrix, while  $B \in \mathbb{R}^{N \times 1}$  and  $C \in \mathbb{R}^{N \times 1}$  represent the projection parameters. The skip connection term  $D \in \mathbb{R}$  can be regarded as a residual connection; thus, the SSM does not necessarily need to include item  $D$ .

### 3.1.2. Discretization

Given that they are continuous-time systems, challenges are faced in integrating SSMs with modern deep-learning frameworks designed for discrete-time operations. They need to be discretized to run efficiently on hardware, allowing deep-learning algorithms like gradient descent and backpropagation to be applied. Discretization also aligns with the discrete nature of real-world data, and it enables the model's complexity to be adjusted by varying the time steps to balance its computation and accuracy.

For the input vector  $x_k \in \mathbb{R}^{L \times D}$  sampled from a signal of length  $L$ , we discretize the continuous parameter matrices  $A$  and  $B$  into  $\bar{A}$  and  $\bar{B}$  by introducing the time scale parameter  $\Delta$  according to the zero-order hold principle [54]. The final results of discretization are as follows:

$$\begin{cases} h_k = \bar{A}h_{k-1} + \bar{B}x_k \\ y_k = Ch_k \\ \bar{A} = e^{\Delta A} \\ \bar{B} = (e^{\Delta A} - I)A^{-1}B \\ \bar{C} = C \end{cases} \quad (2)$$

where  $B, C \in \mathbb{R}^{D \times N}$ , and  $\Delta \in \mathbb{R}^D$ . We can only use first-order Taylor series to approximate  $\bar{B}$ :

$$\bar{B} = (e^{\Delta A} - I)A^{-1}B = (\Delta A)(\Delta A)^{-1}\Delta B = \Delta B \quad (3)$$

Therefore, we can transform the expressions above into the following matrix multiplication form:

$$\begin{cases} \bar{K} = (C\bar{B}, C\bar{A}\bar{B}, \dots, C\bar{A}^{L-1}\bar{B}) \\ y = x\bar{K} \end{cases} \quad (4)$$

where  $K \in \mathbb{R}^L$  represents the convolution kernel, which we can then integrate into the neural network.

### 3.2. Architecture

Figure 1 illustrates the overall architecture of HMCNet. Overall, HMCNet consists of a patch embedding layer, an encoder, a decoder, a projection layer, and skip connections between the encoder and the decoder. We adopted a symmetric structure. The patch embedding layer divides the input image  $x \in \mathbb{R}^{H \times W \times 3}$  into non-overlapping patches of size  $4 \times 4$  and then maps the image dimensions to  $C$ , where we set  $C$  to 32 by default. This process results in the embedded image  $x' \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ . Subsequently, before the patches are input into the encoder for feature extraction, we normalize the input patches using layer normalization [55]. Both the encoder and the decoder consist of three stages, with patch merging operations applied at the end of the first two stages to reduce the height and width of the input features while increasing the number of channels. For example, the output feature map of the first HY-SSM block has dimensions of  $\frac{H}{4} \times \frac{W}{4} \times C$ . After the first patch merging operation, its dimensions change to  $\frac{H}{8} \times \frac{W}{8} \times 2C$ . We use  $[2, 2, 2]$  HY-SSM blocks in the three stages, with the channel count for each stage being  $[C, 2C, 4C]$ .

The structure of the decoder is very similar to that of the encoder. The output of the SSM block from the last stage enters the decoder. While the output of each block undergoes

patch merging in the encoder, in the decoder, a patch expanding operation is applied to the output of each block to halve the number of feature channels and increase its width and height. We use SSM blocks at a ratio of [2, 2, 2], with the channel counts for each stage being  $[4C, 2C, C]$ , respectively. Subsequent to the processing of the decoder, the final projection layer is used to restore the size of the feature map to the same dimensions as those of the input image, resulting in the final segmentation image.

In addition to the overall structure of the encoder and the decoder mentioned above, we also used skip connections between the encoder and the decoder. To maintain a simple network structure and reduce the number of parameters, we used an addition operation.

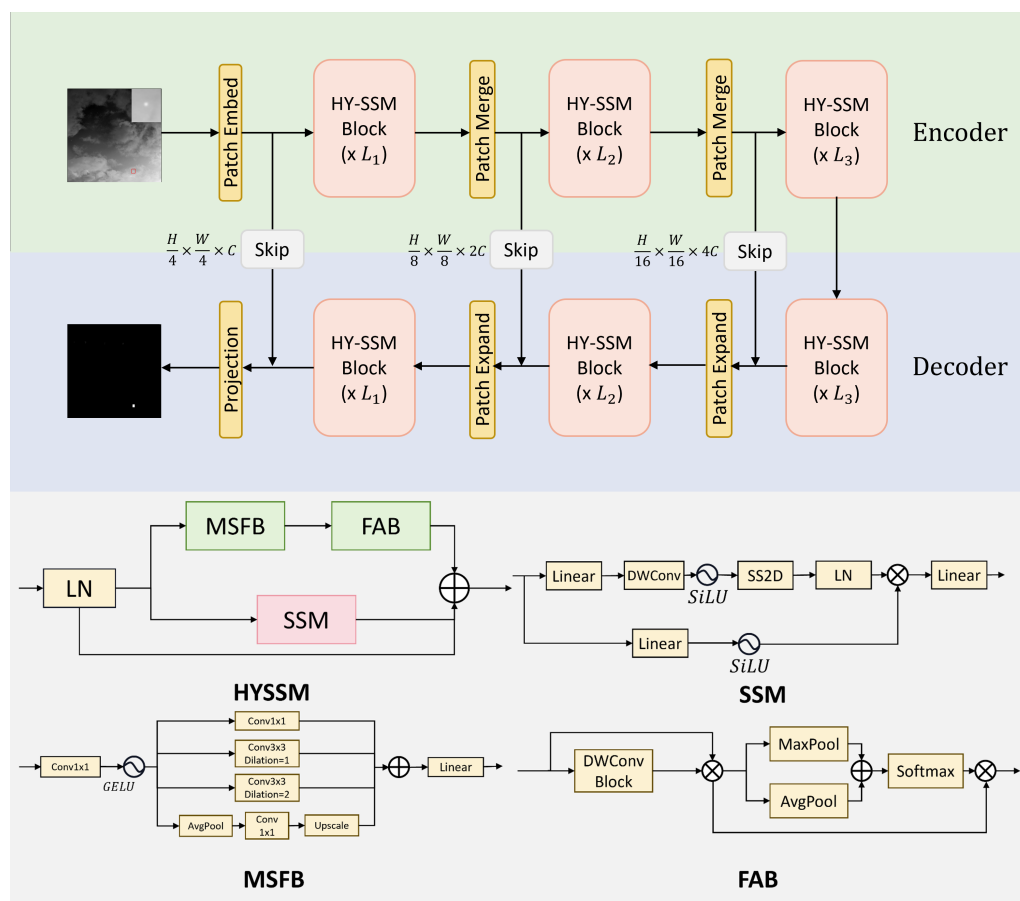


**Figure 1.** Architecture of the model.

### 3.3. The HY-SSM Block

The HY-SSM module is composed of two branch modules: one is the SSM module branch derived from Vmamba [50], and the other is the CNN branch, consisting of a multi-scale feature fusion block (MSFB) and a focus-aware block (FAB). The input feature map first undergoes layer normalization and is then split into the aforementioned two branches for processing, after which the outputs of the two branches are fused with the input feature map.

The CNN branch consists of a multi-scale feature fusion block (MSFB) and a focus-aware block (FAB). The MSFB module consists of a  $1 \times 1$  convolutional layer, a GeLU [56] activation layer, and an Inception module. The Inception module comprises four branches: a  $1 \times 1$  convolutional layer branch, two  $3 \times 3$  convolutional layer branches with dilation coefficients of 1 and 2, and an average pooling branch. After the four branches are added together, they pass through a linear layer to obtain the output of the MSFB module. The output of the FAB module is generated by sequentially merging two attention branches

with the main branch. The first spatial attention branch obtains feature map attention in the width and height dimensions via a depthwise separable convolutional block. The depthwise separable convolutional block consists of a  $3 \times 3$  depthwise separable convolutional layer, a batch normalization layer, and a GeLU activation layer. The second channel attention branch combines average pooling and max pooling of the feature maps, followed by passing them through a SoftMax layer. The two attention branches are sequentially multiplied and merged with the feature map to produce the final output.

The SSM module branch uses SiLU [57] as the activation function. The SSM module is divided into two branches, one of which consists of only two layers—a linear layer and an activation function. The other branch comprises a linear layer, a depthwise separable convolutional layer, and an activation function, after which the input enters the 2D-Selective-Scan (SS2D) module for feature extraction. Subsequently, the extracted features undergo layer normalization and are multiplied element-wise with the output of the first branch to integrate the features from both paths. Finally, the features from the two branches are mixed using a linear layer. The outputs of the two modules are then combined with the original input content to finally generate the output of the HY-SSM module.

The SS2D module is the core module in Mamba and consists of three parts: a scan-expanding operation, an S6 module, and a scan-merging operation. As shown in Figure 2, the scan-expanding operation unfolds the input feature map into independent sequences in four different directions. The scanning strategies in these four directions ensure that each element can interact with the previously scanned samples, thereby integrating information from all the other directions and involving a global receptive field without increasing the linear computational complexity. The S6 module, derived from the Mamba model [53], introduces a selection mechanism on the basis of S4 [30], enhancing its ability to distinguish and retain useful information while filtering out useless information. The S6 module is used to perform feature extraction on the unfolded feature map sequences, obtaining features from the sequences in different directions, and finally, to merge the output sequences through a scan-merging operation to produce an output feature map with the same dimensions as the input feature map.

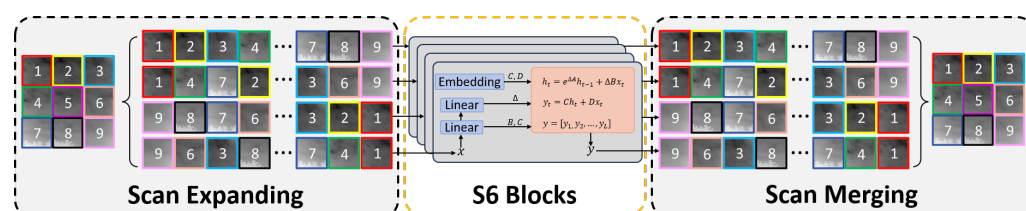


Figure 2. SS 2D module.

Assuming the feature input into the SS2D module is  $z$ , the feature output can be represented as follows:

$$\begin{aligned} z_i &= \text{expand}(z, i) \\ \bar{z}_i &= S6(z_i) \\ \bar{z} &= \text{merge}(\bar{z}_1, \bar{z}_2, \bar{z}_3, \bar{z}_4) \end{aligned} \quad (5)$$

Here,  $i \in \{1, 2, 3, 4\}$  represents the four scanning directions, and  $\text{expand}()$  and  $\text{merge}()$  represent the scan-expanding and scan-merging operations, respectively. A more in-depth explanation of this concept is provided in [50].

### 3.4. The Loss Function

When segmenting small targets in infrared, the targets occupy a very small proportion of the images, while the background takes up a large proportion. The number of negative background samples is several orders of magnitude greater than the number of positive target samples, leading to a severe imbalance between the number of positive and negative samples. This imbalance predisposes the classifier to make predictions for the majority class and ignore the minority class, thereby reducing the overall performance of the model and its ability to make predictions for the minority class. To alleviate this imbalance in the samples, we use the SoftIoU loss function and the focal loss function.

Focal loss (FL) [58] is a loss function used to address imbalances between classes, which are particularly prevalent in object detection and image segmentation tasks. The formula for focal loss is as follows:

$$FL = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (6)$$

where  $p_t$  is the probability predicted by the model;  $\alpha_t$  is the balancing factor for the samples, which adjusts the proportions of different samples; and  $\gamma$  is a tunable parameter for balancing the importance of samples that are easy and hard to classify. When the probability  $p_t$  approaches 1, a sample is considered easy to classify, and it is assigned a smaller weight. Conversely, when  $p_t$  is lower, this indicates a sample is harder to classify, and it is given a higher weight.

SoftIoU loss [59] is a smoothed IoU-based metric designed to optimize the training process more effectively. Compared to traditional IoU loss, SoftIoU loss manages imbalances between classes better and updates the gradients more smoothly, thereby preventing oscillations during training.

$$L_{softiou} = 1 - IoU = 1 - \frac{\sum_i^N P_i \cdot T_i}{\sum_i^N (P_i + T_i - P_i \cdot T_i)} \quad (7)$$

In this formula,  $N$  is the total number of pixels in an image,  $P_i$  represents the probability predicted at the  $i$ -th pixel position, and  $T_i$  denotes the ground-truth label at the  $i$ -th pixel position.

SoftIOU loss is used to stabilize the training process, while focal loss is used to mitigate the problem of class imbalances in samples. The final expression for the loss function is as follows:

$$Loss = FL + L_{SoftIoU} \quad (8)$$

## 4. Experiments

In this section, we will comprehensively introduce the details of the experiment, the dataset, the specifics about the experimental equipment and environment, the evaluation metrics, and the specifics of the design, along with outlining the visual and quantitative results of comparative experiments and the quantitative results of ablation studies.

### 4.1. The Dataset

Our experiments were conducted on the renownedIRSTD-1K dataset [20] and the MISTD dataset we built ourselves.IRSTD-1K consists of 1000 infrared images of real scenes, each with a resolution of  $512 \times 512$  pixels, and all these images are annotated. This dataset features a variety of small targets, including drones and vehicles, within complex backgrounds, such as urban, rural, and river environments. Due to its inclusion of high-resolution, cluttered background noise and diverse types of targets,IRSTD-1K stands out as a benchmark dataset in the field of infrared small-target detection. The MISTD dataset

we built ourselves includes 489 infrared images with a land background captured using the Qilu-1 satellite, and 1173 images with an urban background taken with an infrared mid-wave camera. To simulate small targets, we utilized Gaussian kernels. To introduce more randomness, we applied multiple Gaussian kernels that had undergone random elliptical scaling for random directional translations and concatenations. The peak and standard deviation of the Gaussian kernels were randomly generated. We integrate the NUDT-SIRST [18] public dataset into it, resulting in a hybrid dataset. In this study, we divided the above two datasets into training, validation, and test sets at a ratio of 5:3:2. Statistical information on the dataset is shown in Table 1.

**Table 1.** Dataset statistics.

| Dataset  | Image Number | Resolution | Target Size | Scene                             |
|----------|--------------|------------|-------------|-----------------------------------|
| IRSTD-1K | 1000         | 512 × 512  | 1~1065 pix  | City, Field, Cloud, Sea, Mountain |
| MISTD    | 2753         | 640 × 512  | 1~981 pix   | City, Field, Sea, Cloud           |

#### 4.2. Evaluation Metrics

$P_d$  indicates the model's ability to identify targets correctly, and it is defined as the ratio of the sum of correctly detected targets  $T_{correct}$  to the sum of all targets  $T_{all}$ :

$$P_d = \frac{T_{correct}}{T_{all}} \quad (9)$$

$Fa$  indicates the model's tendency to incorrectly classify negatives as positives, and it is defined as the ratio of false positive pixels  $P_{false}$  to the sum of all pixels  $P_{all}$ :

$$F_a = \frac{P_{false}}{P_{all}} \quad (10)$$

The ROC curve is a tool for evaluating the performance of binary classification models, with the probability of detection ( $P_d$ ) and the false alarm rate ( $Fa$ ) used as the y and x axes here. The curve being closer to the top-left corner represents a better model performance, and the area under the curve (AUC) indicates the overall performance of the model. The ROC curve is helpful for determining the optimal classification threshold and is widely applied in fields such as detection and segmentation.

The IoU (Intersection over the Union) is a commonly used metric for measuring the accuracy of the bounding boxes in object detection algorithms. It is defined as the ratio of the area of the intersection of two bounding boxes to the area of their union. In the field of image segmentation, the IoU is used to evaluate a model's accuracy in classifying each pixel. By calculating the IoU between a model's predicted segmentation result and the ground-truth labels, its segmentation accuracy can be quantified. In this paper, both the Intersection over the Union and the normalized Intersection over the Union [17] are used as evaluation metrics, both of which are pixel-level evaluation metrics.

The IoU is a metric used to measure the degree of overlap between two sets, typically between two regions (i.e., masks in this work). It is defined as follows:

$$\text{IoU} = \frac{TP}{T + P - TP} \quad (11)$$

where TP, T, and P denote the true positive pixels, true pixels, and positive pixels, respectively.

The nIoU was specifically tailored to the SIRST dataset to provide a more equitable metric that bridges the gap between model-driven and data-driven approaches. It is defined as follows:

$$\text{nIoU} = \frac{1}{N} \sum_i^N \frac{TP[i]}{T[i] + P[i] - TP[i]} \quad (12)$$

where  $N$  is the total number of targets.

Floating-point operations (FLOPs) and parameters (Params) are metrics used to assess the complexity of deep-learning models. FLOPs indicates the number of floating-point operations a model performs per second, while Params indicates the number of parameters the model needs to learn. These metrics help us to understand the computational demands and scale of a model, thereby aiding in model selection, optimization, and deployment. The magnitude of FLOPs reflects the complexity of a model; models with a higher number of FLOPs are more computationally complex. The number of Params reflects the size of a model, and smaller models involve lower computational burdens.

#### 4.3. Implementation Details

We conducted sufficient experiments on the model proposed in this paper, including comparative experiments using the IRSTD-1k dataset and our own MISTD dataset, as well as ablation studies. The details of our training equipment are as follows: an NVIDIA RTX 2080 GPU (8 GB), 32 GB of RAM, an Intel Core i7-10700K CPU (3.8 GHz), the WSL operating system (Windows Subsystem for Linux), the deep-learning framework PyTorch 2.2, and Python 3.10. The experimental setup is as follows: images were uniformly resized to  $256 \times 256$ , the optimizer used was Adaptive Gradient (AdaGrad), the weight initialization method was Xavier initialization, the learning rate was set to 0.004, the number of training epochs was 500, the batch size was 8, and the loss functions used were focal and SoftIoU loss. Table 2 shows the details of the experimental configuration of the traditional methods in the comparative models.

**Table 2.** Configurations of model-based methods.

| Method | Configuration                                                                                                   |
|--------|-----------------------------------------------------------------------------------------------------------------|
| MPCM   | kernel size: 3, 5, 7, 9                                                                                         |
| RLCM   | kernel size: 3, 5, 7, 9                                                                                         |
| PSTNN  | patch size: $40 \times 40$ , step: 40, $\epsilon = 10^{-7}$ , $\lambda = 0.7 / (\sqrt{\min(n1, n2)} \times n3)$ |

#### 4.4. Comparison with Various Methods

The quantitative results of the comparison given in Table 3 show that our method achieved the best performance on the IRSTD-1k and MISTD datasets while also achieving a balance between accuracy and the number of parameters. Traditional-model-based methods tend to perform poorly when dealing with complex backgrounds and datasets containing varying sizes and shapes, whereas data-driven deep-learning methods generally adapt well to the sizes of the targets and can effectively suppress background noise. Compared to other deep-learning approaches, our model has an advantage in capturing very small targets due to the combination of the powerful attention mechanism of the Mamba module and the capabilities of CNNs in local feature integration. Figure 3 shows an intuitive comparison of model performance.

**Table 3.** Comparison of using different methods on IRSTD-1k and MISTD in terms of *IoU* (%), *nIoU* (%), *Pd* (%), and *Fa* ( $10^{-5}$ ).

| Method                | IRSTD-1k    |             |             |                  | MISTD       |             |             |                  |
|-----------------------|-------------|-------------|-------------|------------------|-------------|-------------|-------------|------------------|
|                       | IoU (%)     | nIoU (%)    | Pd (%)      | Fa ( $10^{-5}$ ) | IoU (%)     | nIoU (%)    | Pd (%)      | Fa ( $10^{-5}$ ) |
| Model-based           | -           | -           | -           | -                | -           | -           | -           | -                |
| RLCM                  | 32.0        | 25.2        | 77.6        | 46.3             | 31.2        | 33.3        | 65.9        | 46.9             |
| MPCM                  | 24.8        | 22.7        | 80.9        | 37.1             | 24.9        | 21.8        | 66.5        | 33.4             |
| PSTNN                 | 28.1        | 35.6        | 85.2        | 39.6             | 40.2        | 42.5        | 78.9        | 24.1             |
| CNN-based             | -           | -           | -           | -                | -           | -           | -           | -                |
| MDFA                  | 48.3        | 46.8        | 83.7        | 38.4             | 50.1        | 48.5        | 84.4        | 23.1             |
| DNA <sub>net</sub>    | 63.4        | 64.2        | 93.5        | 4.31             | 69.6        | 70.5        | 92.7        | 5.23             |
| AMFUnet               | 54.1        | 50.9        | 89.1        | 10.3             | 62.8        | 59.6        | 89.7        | 10.03            |
| UIUnet                | 67.4        | 66.7        | <b>95.4</b> | 3.77             | 69.8        | 68.7        | 91.4        | 5.42             |
| RDIAN                 | 62.9        | 63.3        | 92.2        | 3.05             | 65.9        | 66.1        | 88.6        | 6.98             |
| CNN-Transformer-based | -           | -           | -           | -                | -           | -           | -           | -                |
| IAA <sub>net</sub>    | 61.5        | 59.4        | 86.2        | 16.9             | 57.2        | 57.9        | 90.7        | 9.88             |
| CNN-Mamba-based       | -           | -           | -           | -                | -           | -           | -           | -                |
| RDIAN+SSM             | 63.2        | 63.5        | 92.3        | 3.03             | 66.3        | 66.2        | 89.0        | 6.79             |
| Ours                  | <b>68.0</b> | <b>67.7</b> | <b>94.4</b> | <b>2.96</b>      | <b>72.6</b> | <b>71.2</b> | <b>93.4</b> | <b>5.22</b>      |

Note: Bold values indicate the best performance.

To test whether the *IoU* values of our proposed method are superior to those of other high-performing methods, we designed significance tests for three other high-performing methods, as shown in Table 4. We applied the Student's paired two-sample t-test, where each experiment fixed the same random seed to initialize different model parameters, retrained the models, and repeated the experiments 10 times to obtain 10 sets of *IoU* values. Subsequently, we conducted the Student's t-test on these 10 sets of *IoU* values to examine whether the differences in *IoU* values were significant. The significance level was set at  $\alpha = 0.05$ . The results of the significance tests indicate that our method outperforms the other methods in terms of *IoU* values, with p-values being less than 0.05. This demonstrates that the superiority of our method over the other comparison methods in terms of *IoU* values is significant.

**Table 4.** Significance test of different method of *IoU*(%) in IRSTD-1k.

| Method | Avg IoU(%) | Var IoU(%) | df | t Stat   | P(T≤t)<br>Single-Tailed | t Single-Tailed<br>Crucial Value | P(T≤t)<br>Two-Tailed  | t Two-Tailed<br>Crucial Value |
|--------|------------|------------|----|----------|-------------------------|----------------------------------|-----------------------|-------------------------------|
| UIUnet | 67.41      | 0.01433    | 9  | −13.0486 | $1.88 \times 10^{-7}$   | 1.833                            | $3.76 \times 10^{-7}$ | 2.2621                        |
| Ours   | 68.01      | 0.01156    | -  | -        | -                       | -                                | -                     | -                             |

A visual comparison of using our proposed method and other methods on the IRSTD-1k and MISTD datasets is presented in Figures 4–10. In these figures, we use red square boxes to frame detected targets, blue circles to frame missed targets, and yellow boxes to frame false detections. We also show a local amplification of the detection target. The traditional methods (RLCM, MPCM, and PSTNN) can detect brighter targets, but brighter noise from non-targets will also be detected, resulting in a higher false detection rate. For weak targets, the missed detection rate is too high. This is because the detection mechanism in the traditional methods is based on the difference in the intensity between the target and the background without considering information on the target's shape and structure information; thus, the adaptability of this kind of scheme is poor. The scheme of the CNN is obviously superior to that of the traditional methods in dealing with the background. It adapts well to the background and provides good detection results in most scenarios.

However, in the presence of complex backgrounds, missed detections and false detections still arise. The lightweight RDIAN model performs well in scenes with simple backgrounds but performs poorly where complex backgrounds and weak targets are concerned. Similar phenomena apply in several of the other models. In terms of retaining the shape features of targets, our method also exhibits a better performance than that of the other models, as well as better adaptability to complex backgrounds. As shown in the figures, our method yields good detection results in different scenarios. This is due to its combined ability to model the global context and capture local features, suppressing complex backgrounds, extracting targets from these complex backgrounds, and preserving the targets' shape information well.

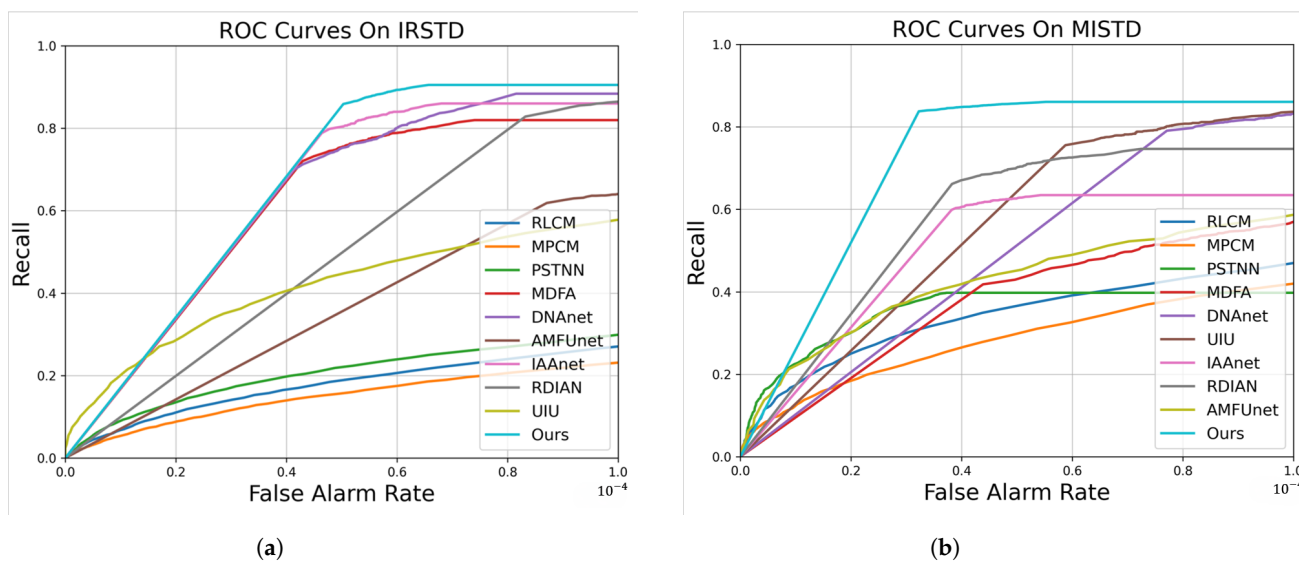


Figure 3. (a) ROC curves for IRSTD-1k. (b) ROC curves for MISTD.

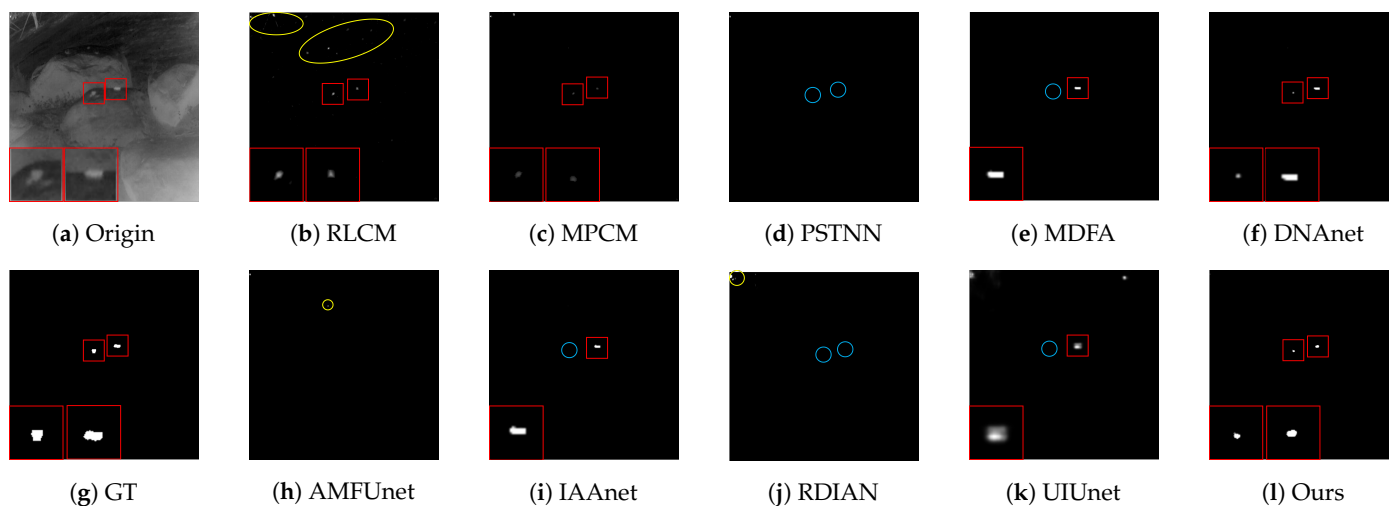


Figure 4. Visual comparison of different methods in IRSTD-1k dataset scene 1.

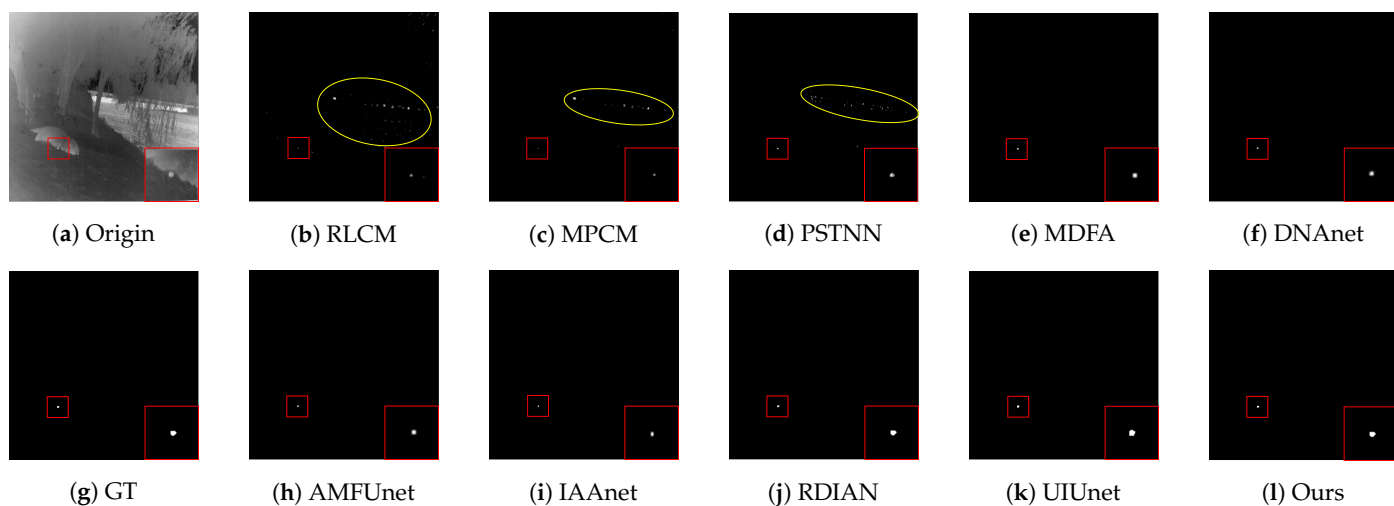


**Figure 5.** Visual comparison of different methods inIRSTD-1k dataset scene 2.

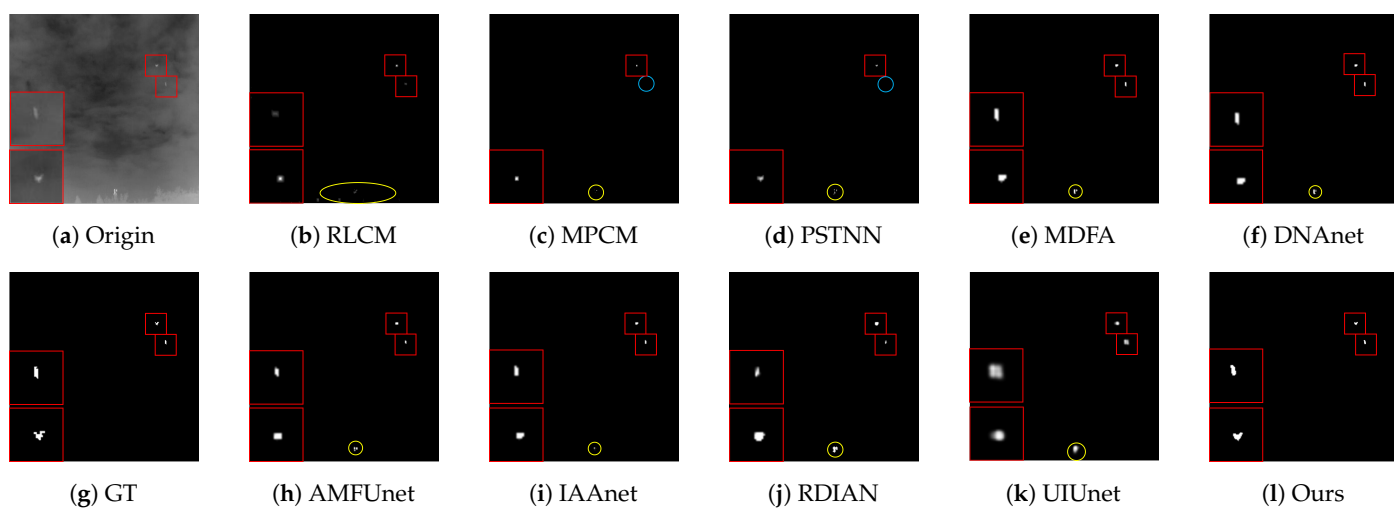


**Figure 6.** Visual comparison of different methods inIRSTD-1k dataset scene 3.

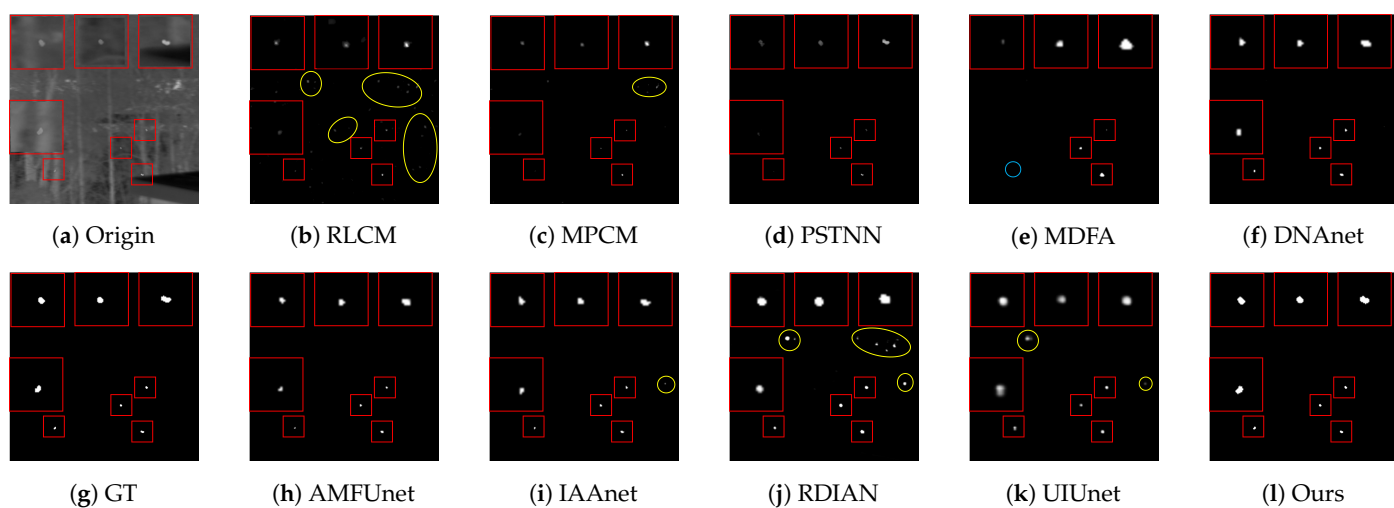


**Figure 7.** Visual comparison of different methods inMISTD dataset scene 1.

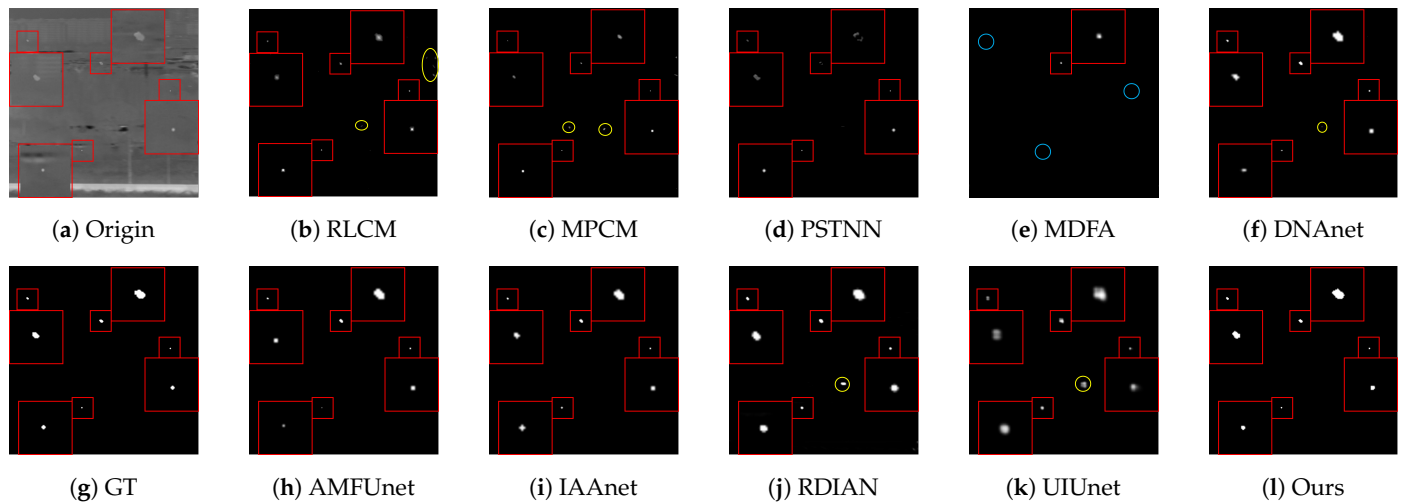


**Figure 8.** Visual comparison of different methods in MISTD dataset scene 2.

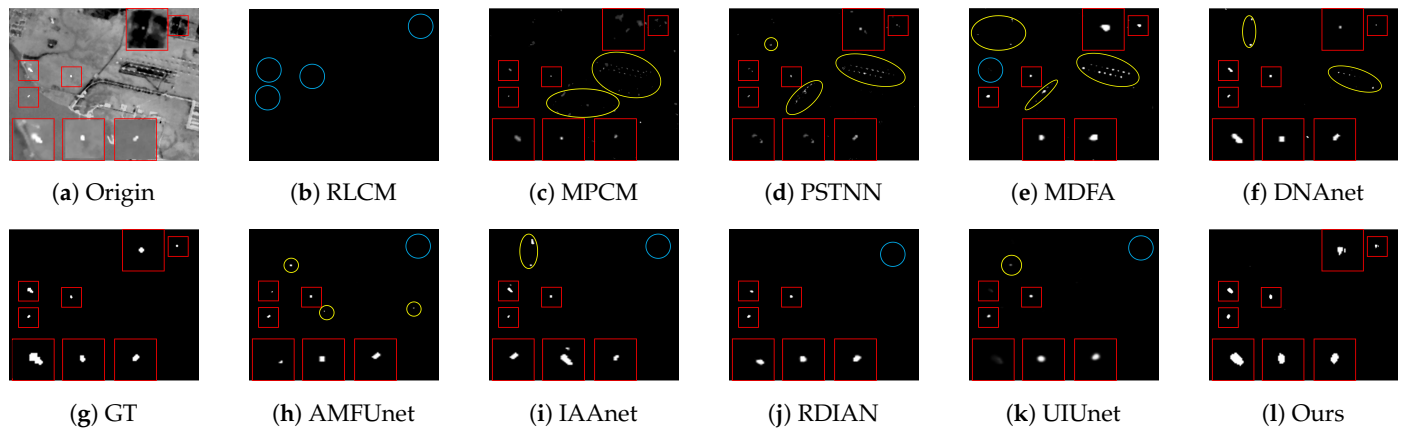


**Figure 9.** Visual comparison of different methods in MISTD dataset scene 3.

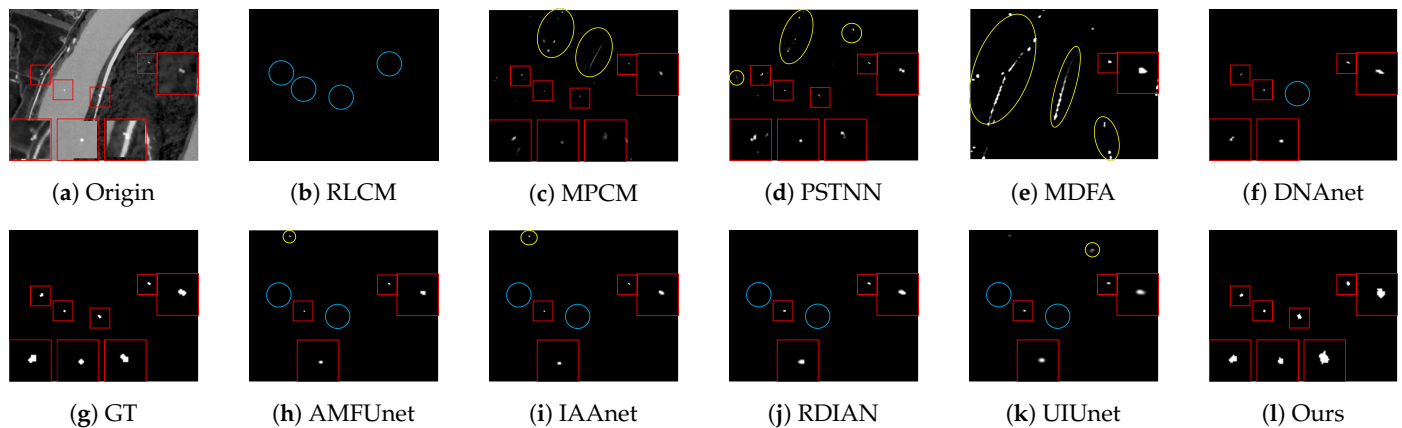


**Figure 10.** Visual comparison of different methods in MISTD dataset scene 4.

Table 5 shows the computational performance of the deep-learning-based models on the IRSTD-1k and MISTD datasets. It can be seen from the table that most of the models involve a large amount of calculation, making it difficult for them to meet the needs of real-time processing. The UIUnet model represents a particularly extreme example of this in that both its accuracy and number of parameters are high. The accuracy of the RDIAN lightweight model is lower than that of the other heavyweight models, but its real-time performance is very good. Our model achieves a good balance between the amount of calculation required and its accuracy, maintaining high accuracy and involving a lower amount of calculation.

**Table 5.** Computational performance of different deep-learning methods.

| Method      | Params (M) | FLOPs (G) | IoU (%) | FPS   |
|-------------|------------|-----------|---------|-------|
| MDFA        | 15.56      | 61.93     | 48.3    | 8.4   |
| AMFUnet     | 0.5        | 23.19     | 54.1    | 104.2 |
| DNAnet      | 4.7        | 56.71     | 63.4    | 65.7  |
| UIUnet      | 50.5       | 65.83     | 67.4    | 50.7  |
| IAAnet      | 14.1       | 18.16     | 61.5    | 43.9  |
| RDIAN       | 0.22       | 4.45      | 62.9    | 186.1 |
| RDIAN + SSM | 0.51       | 4.98      | 63.2    | 136.5 |
| Ours        | 0.9        | 5.82      | 68.0    | 148.3 |

#### 4.5. The Ablation Study

To evaluate the contribution of each component to our model, we conducted a series of ablation studies. After systematically removing or modifying specific parts of the model, we then observed the resulting changes in its performance. This approach helped us understand the significance of each component and its impact on the model's overall effectiveness.

##### 4.5.1. The Ablation Study on the SSM Branch

To verify the contribution of the SSM branch to the detection accuracy, we conducted ablation experiments on the SSM branch separately using the two datasets IRSTD-1k and MISTD. The results in Table 6 show that removing the SSM module had a significant negative impact on the model's accuracy, with its IoU and nIoU decreasing by 12.0% and 16.0%, respectively, for the IRSTD-1k dataset and Pd decreasing by 8.8% and Fa increasing by  $10.74 \times 10^{-5}$ . For the MISTD dataset, its IoU and nIoU decreased by 9.1% and 6.8%, respectively; Pd decreased by 6.9%; and Fa increased by  $5.37 \times 10^{-5}$ . These results demonstrate the SSM module's ability to handle complex backgrounds and noise, as well as the effectiveness of its global attention mechanism in improving detection accuracy.

**Table 6.** Results of ablation study on SSM and CNN modules.

| Method           | IRSTD-1k |          |        |                  | MISTD   |          |        |                  |
|------------------|----------|----------|--------|------------------|---------|----------|--------|------------------|
|                  | IoU (%)  | nIoU (%) | Pd (%) | Fa ( $10^{-5}$ ) | IoU (%) | nIoU (%) | Pd (%) | Fa ( $10^{-5}$ ) |
| FAB + MSFB       | 56.2     | 51.7     | 85.6   | 13.7             | 63.5    | 64.4     | 87.9   | 9.68             |
| SSM + FAB        | 61.3     | 58.6     | 86.4   | 16.5             | 62.7    | 64.3     | 87.4   | 11.3             |
| SSM + MSFB       | 64.5     | 62.1     | 87.2   | 12.9             | 68.3    | 68.1     | 91.0   | 8.56             |
| SSM + FAB + MSFB | 68.0     | 67.7     | 94.4   | 2.96             | 72.6    | 71.2     | 93.4   | 5.22             |

##### 4.5.2. The Ablation Study on the CNN Branch

The CNN branch is divided into two modules, namely the FAB and the MSFB. We conducted ablation studies on these two modules to explore their impact on the model's detection accuracy. The results in Table 6 show that when using the FAB module alone, the model's IoU and nIoU on the IRSTD-1k dataset decreased by 6.7% and 9.1%, respectively; Pd decreased by 8.0%; and Fa increased by  $13.54 \times 10^{-5}$ . When using the MSFB module alone, the model's IoU and nIoU on the IRSTD-1k dataset decreased by 3.4% and 5.6%, respectively; Pd decreased by 7.2%; and Fa increased by  $9.94 \times 10^{-5}$ . Similar trends were observed for the MISTD dataset. These results indicate that the capabilities of the FAB and MSFB modules in local feature extraction contribute to the model's improvement in detection accuracy.

#### 4.5.3. The Ablation Study on Loss Function

We conducted an ablation experiment to compare the selection of loss functions to verify the contribution of the loss function used in this method to the training effect. We selected two loss functions, Focal loss and SoftIoU loss, respectively, and obtained the comparison results on the data set, as shown in Table 7. As can be seen from the table, each evaluation indicator has a certain degree of loss. Therefore, we conclude that using two loss functions alone is not as good as using two loss functions at the same time for the final training results. This result verifies the contribution of the two loss functions, Focal loss and SoftIoU loss, to the training results. Their combination in use can achieve the best training accuracy.

**Table 7.** Results of ablation study on Loss Functions.

| Loss            | IRSTD-1k |          |        |                  | MISTD   |          |        |                  |
|-----------------|----------|----------|--------|------------------|---------|----------|--------|------------------|
|                 | IoU (%)  | nIoU (%) | Pd (%) | Fa ( $10^{-5}$ ) | IoU (%) | nIoU (%) | Pd (%) | Fa ( $10^{-5}$ ) |
| Focal Loss      | 66.1     | 66.5     | 89.7   | 5.84             | 70.8    | 68.4     | 91.7   | 6.78             |
| SoftIoU Loss    | 65.3     | 66.0     | 88.9   | 6.31             | 68.6    | 68.2     | 91.1   | 6.56             |
| Focal + SoftIoU | 68.0     | 67.7     | 94.4   | 2.96             | 72.6    | 71.2     | 93.4   | 5.22             |

## 5. Conclusions

Infrared small weak target detection is applied extensively in fields such as security monitoring, with complex backgrounds and weak target signals constituting the main challenges in this domain. The traditional-model-based approaches have a limited capacity to model these targets, while deep-learning methods, although they have achieved significant progress, still face issues like the disappearance of features and high computational complexity. In this paper, we propose a meticulously designed architecture for the detection of small targets, with which we aim to alleviate the problems of feature disappearance for weak infrared targets and the high computational complexity of global attention mechanisms. We used a hybrid module combining a CNN and a state space model (SSM) as the feature-extraction component in the UNet backbone. The CNN branch employed a multi-scale fusion module and a feature-focused module, whereas the SSM branch was used to model the global context, powering the model with a global attention mechanism and mitigating the issue of low-level feature disappearance in deep networks. This model also demonstrated a superior computational efficiency compared to that of the mainstream methods, achieving a suitable balance between its computational volume and accuracy. Through comparative analysis with the current state-of-the-art methods on the public dataset IRSTD-1k and our custom dataset MISTD, we showcased its outstanding performance. We also designed comprehensive ablation experiments that verified the role of each component, proving the effectiveness of our method for segmenting weak and small targets in infrared and representing a notable advancement in this field.

**Author Contributions:** Conceptualization, B.L.; methodology, B.L. and P.R.; validation, B.L.; writing, B.L. and Y.S.; editing, X.C. and P.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the Talent Plan of Shanghai Branch, the Chinese Academy of Sciences, Grant No. CASSHB-QNPD-2023-007.

**Data Availability Statement:** The IRSTD-1k dataset supporting this research can be obtained at the following website : <https://github.com/RuiZhang97/ISNet> (accessed on 20 March 2022). The NUDT-SIRST dataset supporting this research can be obtained at the following website: <https://github.com/YeRen123455/Infrared-Small-Target-Detection> (accessed on 16 August 2022).

**Acknowledgments:** The authors would like to thank Z.M. and B.Y. for providing the open-source data.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Teutsch, M.; Krüger, W. Classification of small boats in infrared images for maritime surveillance. In Proceedings of the 2010 International WaterSide Security Conference, Carrara, Italy, 3–5 November 2010; pp. 1–7.
- Zhao, M.; Li, W.; Li, L.; Hu, J.; Ma, P.; Tao, R. Single-frame infrared small-target detection: A survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 87–119. [\[CrossRef\]](#)
- Rawat, S.S.; Verma, S.K.; Kumar, Y. Review on recent development in infrared small target detection algorithms. *Procedia Comput. Sci.* **2020**, *167*, 2496–2505. [\[CrossRef\]](#)
- Cheng, Y.; Lai, X.; Xia, Y.; Zhou, J. Infrared Dim Small Target Detection Networks: A Review. *Sensors* **2024**, *24*, 3885. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tukey, J. Nonlinear (nonsuperposable) methods for smoothing data. In Proceedings of the Congress Record EASCON, Washington, DC, USA, 7–9 October 1974; p. 673.
- Deshpande, S.D.; Er, M.H.; Venkateswarlu, R.; Chan, P. Max-mean and max-median filters for detection of small targets. In Proceedings of the Signal and Data Processing of Small Targets 1999, Orlando, FL, USA, 20–22 July 1999; Volume 3809, pp. 74–83.
- Rivest, J.F.; Fortin, R. Detection of dim targets in digital infrared imagery by morphological image processing. *Opt. Eng.* **1996**, *35*, 1886–1893. [\[CrossRef\]](#)
- Tom, V.T.; Peli, T.; Leung, M.; Bondaryk, J.E. Morphology-based algorithm for point target detection in infrared backgrounds. In Proceedings of the Signal and Data Processing of Small Targets 1993, Orlando, FL, USA, 12–14 April 1993; Volume 1954, pp. 2–11.
- Zeng, M.; Li, J.; Peng, Z. The design of top-hat morphological filter and application to infrared target detection. *Infrared Phys. Technol.* **2006**, *48*, 67–76. [\[CrossRef\]](#)
- Chen, C.P.; Li, H.; Wei, Y.; Xia, T.; Tang, Y.Y. A local contrast method for small infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 574–581. [\[CrossRef\]](#)
- Lv, P.; Sun, S.; Lin, C.; Liu, G. A method for weak target detection based on human visual contrast mechanism. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 261–265. [\[CrossRef\]](#)
- Han, J.; Xu, Q.; Moradi, S.; Fang, H.; Yuan, X.; Qi, Z.; Wan, J. A Ratio-Difference Local Feature Contrast Method for Infrared Small Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 7506705. [\[CrossRef\]](#)
- Han, J.; Liang, K.; Zhou, B.; Zhu, X.; Zhao, J.; Zhao, L. Infrared small target detection utilizing the multiscale relative local contrast measure. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 612–616. [\[CrossRef\]](#)
- Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [\[CrossRef\]](#)
- Wang, H.; Zhou, L.; Wang, L. 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, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8509–8518.
- Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric contextual modulation for infrared small target detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 950–959.
- Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional local contrast networks for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [\[CrossRef\]](#)
- Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense nested attention network for infrared small target detection. *IEEE Trans. Image Process.* **2022**, *32*, 1745–1758. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wu, X.; Hong, D.; Chanussot, J. UIU-Net: U-Net in U-Net for Infrared Small Object Detection. *IEEE Trans. Image Process.* **2023**, *32*, 364–376. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape matters for infrared small target detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 877–886.
- Sun, H.; Bai, J.; Yang, F.; Bai, X. Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5000513. [\[CrossRef\]](#)
- Dosovitskiy, A. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 11–15 June 2021; pp. 10012–10022.
- He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [\[CrossRef\]](#)

25. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
26. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. 3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers. *arXiv* **2023**, arXiv:2310.07781.
27. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
28. Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.R.; Xu, D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 272–284.
29. Wang, K.; Du, S.; Liu, C.; Cao, Z. Interior attention-aware network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5002013. [[CrossRef](#)]
30. Gu, A.; Goel, K.; Ré, C. Efficiently Modeling Long Sequences with Structured State Spaces. *arXiv* **2022**, arXiv:cs.LG/2111.00396.
31. Nguyen, E.; Goel, K.; Gu, A.; Downs, G.; Shah, P.; Dao, T.; Baccus, S.; Ré, C. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2846–2861.
32. Islam, M.M.; Bertasius, G. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 87–104.
33. Chen, T.; Ye, Z.; Tan, Z.; Gong, T.; Wu, Y.; Chu, Q.; Liu, B.; Yu, N.; Ye, J. Mim-istd: Mamba-in-mamba for efficient infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5007613. [[CrossRef](#)]
34. Han, J.; Ma, Y.; Zhou, B.; Fan, F.; Liang, K.; Fang, Y. A robust infrared small target detection algorithm based on human visual system. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 2168–2172.
35. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A.G. Infrared patch-image model for small target detection in a single image. *IEEE Trans. Image Process.* **2013**, *22*, 4996–5009. [[CrossRef](#)]
36. Zhou, F.; Wu, Y.; Dai, Y.; Wang, P. Detection of small target using Schatten  $1/2$  quasi-norm regularization with reweighted sparse enhancement in complex infrared scenes. *Remote Sens.* **2019**, *11*, 2058. [[CrossRef](#)]
37. Li, W.; Zhao, M.; Deng, X.; Li, L.; Li, L.; Zhang, W. Infrared small target detection using local and nonlocal spatial information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3677–3689. [[CrossRef](#)]
38. Zhang, X.; Ding, Q.; Luo, H.; Hui, B.; Chang, Z.; Zhang, J. Infrared small target detection based on an image-patch tensor model. *Infrared Phys. Technol.* **2019**, *99*, 55–63. [[CrossRef](#)]
39. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [[CrossRef](#)]
40. Zhang, W.; Cong, M.; Wang, L. Algorithms for optical weak small targets detection and tracking. In Proceedings of the International Conference on Neural Networks and Signal Processing, Nanjing, China, 14–17 December 2003; Volume 1, pp. 643–647.
41. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
42. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018.
43. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–9 May 2020; pp. 1055–1059.
44. Chung, W.Y.; Lee, I.H.; Park, C.G. Lightweight infrared small target detection network using full-scale skip connection U-Net. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 7000705. [[CrossRef](#)]
45. Kim, J.H.; Hwang, Y. GAN-Based Synthetic Data Augmentation for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5002512. [[CrossRef](#)]
46. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
47. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
48. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

49. Liu, F.; Gao, C.; Chen, F.; Meng, D.; Zuo, W.; Gao, X. Infrared small and dim target detection with transformer under complex backgrounds. *IEEE Trans. Image Process.* **2023**, *32*, 5921–5932. [[CrossRef](#)] [[PubMed](#)]
50. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. VMamba: Visual State Space Model. *arXiv* **2024**, arXiv:2401.10166.
51. Ma, J.; Li, F.; Wang, B. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. *arXiv* **2024**, arXiv:2401.04722.
52. Wang, Z.; Zheng, J.Q.; Zhang, Y.; Cui, G.; Li, L. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv* **2024**, arXiv:2402.05079.
53. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
54. Gupta, A.; Gu, A.; Berant, J. Diagonal state spaces are as effective as structured state spaces. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22982–22994.
55. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
56. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2023**, arXiv:1606.08415.
57. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *arXiv* **2017**, arXiv:1702.03118. [[CrossRef](#)] [[PubMed](#)]
58. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
59. Rahman, M.A.; Wang, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In Proceedings of the International Symposium on Visual Computing, Las Vegas, NV, USA, 12–14 December 2016; pp. 234–244.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.