



## Article

# MM-IRSTD: Conv Self-Attention-Based Multi-Modal Small and Dim Target Detection in Infrared Dual-Band Images

Junyan Yang <sup>1,2</sup> , Zhihui Ye <sup>1</sup> , Jian Lin <sup>1</sup> , Dongfang Chen <sup>2</sup> , Lingbian Du <sup>2</sup> and Shaoyi Li <sup>1,\*</sup>

<sup>1</sup> School of Astronautics, Northwestern Polytechnical University, Xi'an 710072, China; yx949456484@mail.nwpu.edu.cn (J.Y.); yezhihui@mail.nwpu.edu.cn (Z.Y.); linaom1214@mail.nwpu.edu.cn (J.L.)

<sup>2</sup> Shanghai Academy of Spaceflight Technology, Shanghai 201109, China; nwpu\_lj@mail.nwpu.edu.cn (D.C.); nwpu\_lys@mail.nwpu.edu.cn (L.D.)

\* Correspondence: 2015030129@nwpu.edu.cn

**Abstract:** Infrared multi-band small and dim target detection is an important research direction in the fields of modern remote sensing and military surveillance. However, achieving high-precision detection remains challenging due to the small scale, low contrast of small and dim targets, and their susceptibility to complex background interference. This paper innovatively proposes a dual-band infrared small and dim target detection method (MM-IRSTD). In this framework, we integrate a convolutional self-attention mechanism module and a self-distillation mechanism to achieve end-to-end dual-band infrared small and dim target detection. The Conv-Based Self-Attention module consists of a convolutional self-attention mechanism and a multilayer perceptron, effectively extracting and integrating input features, thereby enhancing the performance and expressive capability of the model. Additionally, this module incorporates a dynamic weight mechanism to achieve adaptive feature fusion, significantly reducing computational complexity and enhancing the model's global perception capability. During model training, we use a spatial and channel similarity self-distillation mechanism to drive model updates, addressing the similarity discrepancy between long-wave and mid-wave image features extracted through deep learning, thus improving the model's performance and generalization capability. Furthermore, to better learn and detect edge features in images, this paper designs an edge extraction method based on Sobel. Finally, comparative experiments and ablation studies validate the advancement and effectiveness of our proposed method.

**Keywords:** dual-band infrared; target detection; attention mechanism; feature fusion; lightweight network; Sobel; KL divergence



**Citation:** Yang, J.; Ye, Z.; Lin, J.; Chen, D.; Du, L.; Li, S. MM-IRSTD: Conv Self-Attention-Based Multi-Modal Small and Dim Target Detection in Infrared Dual-Band Images. *Remote Sens.* **2024**, *16*, 3937. <https://doi.org/10.3390/rs16213937>

Academic Editors: Shou Feng, Nan Su, Chunhui Zhao, Yiming Yan, Yunsong Li, Bobo Xi and Hao Li

Received: 19 August 2024

Revised: 12 October 2024

Accepted: 16 October 2024

Published: 23 October 2024



**Copyright:** © 2024 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

Infrared detection technology is widely used in various detection fields due to its passive detection method. In practical detection tasks, the infrared images of targets often appear weak and small, and various natural and artificial interferences further increase the difficulty of detection and recognition [1–4]. To fully capture the infrared image information of targets and enhance the ability to resist complex background interference, dual-band infrared target detection technology has become a current research hotspot. Compared to single-band target detection, dual-band target detection has the following significant advantages: (1) Dual-band target detection can fully utilize the redundancy and complementarity of information from different bands, thereby more effectively distinguishing between background and target, reducing false alarm rates, and improving target detection rates; (2) By integrating image information from dual bands, more detailed features of the target can be obtained, thereby enhancing the system's anti-interference capability.

Traditional methods for infrared small and dim target detection can be mainly divided into filter-based methods [5], human visual system-based methods [6] and low-rank sparse recovery-based methods [7]. However, traditional methods require a significant amount

of time for manual tuning and have limited target detection capabilities in complex environments, with poor generalization and robustness. In recent years, the development of artificial intelligence theories and computer hardware has led to the gradual replacement of traditional infrared small and dim target detection methods with deep learning-based methods [8]. Deep learning-based target detection methods are primarily categorized into two-stage methods and one-stage methods. Two-stage methods, represented by the Regional Convolutional Neural Network (R-CNN) series [9–13], offer high detection accuracy but consume a lot of memory and have slow computation speeds. One-stage methods, including the YOLO series [14–24] and the SSD series [25,26], have lower computational requirements and higher real-time performance but cannot effectively capture the detailed information of small targets, resulting in lower detection accuracy.

The attention mechanism, by simulating human attention, allows models to focus more on important parts while ignoring irrelevant ones when processing information. In target detection tasks, the attention mechanism enables models to more accurately focus on target regions, thereby improving the accuracy and robustness of target detection. Transformer is a neural network architecture based on self-attention mechanisms. In 2020, Carion et al. [27] developed a new object detection framework called DETR (Detection Transformer), marking the first application of Transformer to the field of target detection. Since then, with the advent of models such as Deformable DETR [28], TSP [29], UP-DETR [30], ACT [31], Transformer-based models have been widely applied in target detection tasks. Compared to CNNs, the self-attention mechanism of Transformers is not limited by local interactions, allowing them to capture long-range dependencies and perform parallel computations. This enables them to learn the most suitable inductive biases for different task objectives, demonstrating excellent performance in target detection tasks.

In this paper, we innovatively propose an attention mechanism-based framework (MM-IRSTD: Conv Self Attention-Based Multi-Modal Small and Dim Target Detection in Infrared Dual-Band Images), which deeply integrates the convolutional self-attention module with a self-distillation learning mechanism, specifically designed to achieve efficient and accurate end-to-end dual-band infrared small and dim target detection. Specifically, the Conv-Based Self-Attention module we developed combines the convolutional self-attention mechanism with high feature extraction capabilities and the multilayer perceptron (MLP) with complex feature processing abilities. This combination significantly enhances the model's capacity for effective extraction and deep fusion of input infrared image features, thereby improving the overall performance of the model. To further advance model optimization and generalization, we introduced a spatial and channel similarity self-distillation mechanism during the training phase based on KL (Kullback–Leibler) divergence. This module drives model updates by computing and minimizing the similarity differences between long-wave and mid-wave infrared images at the deep feature level, effectively promoting the model's learning and understanding of the intrinsic relationships between features from different bands. This process not only enhances the model's adaptability to complex infrared scenes but also significantly improves detection accuracy and robustness. Finally, through rigorous comparative and ablation experiments, we thoroughly validated the superior performance and significant advantages of the MM-IRSTD in dual-band infrared small and dim target detection tasks, demonstrating its effectiveness and innovation in enhancing detection efficiency and accuracy.

The key achievements and contributions in this paper are as follows:

- We proposed an end-to-end dual-band infrared small and dim target detection model framework that combines the Conv-Based Self-Attention module with the self-distillation mechanism.
- We introduced a Conv-Based Self-Attention module that incorporates dynamic weights to achieve adaptive feature fusion, uses convolutional operations to reduce the method's computational complexity, and enhances the model's global perception capabilities.
- We proposed a KL divergence-based self-distillation mechanism that constrains multi-band feature information from the spatial and channel dimensions of the feature maps.

The self-distillation mechanism is used only during the model's training process, ensuring efficient inference.

- Experimental results show that compared to SOTA approaches, MM-IRSTD improves F-score and Kappa by 3.13% and 7.73%, respectively, in dual-band infrared small and dim target detection tasks.

## 2. Related Work

### 2.1. Target Detection Based on Multiband Image

The current research on multiband target detection mainly focuses on infrared and visible light fusion, as well as infrared multiband detection. Kou et al. [32] proposed an efficient spectral signal recognition method based on an optimized AdaBoost algorithm to address the challenge of single-band infrared point target recognition. This method improves recognition accuracy and speed, reduces noise impact, and avoids overfitting. Liu et al. [33] introduced a low-slow-small target recognition method based on the fusion of multi-band images and spectra. This approach utilizes wide-spectrum infrared data, integrating global and local features, and combines multi-band images with spectral information. It effectively enhances the recognition accuracy of small targets while improving anti-interference capabilities, resulting in more accurate airborne target recognition. Guo et al. [34] proposed a novel infrared differential detection (IDD) method to obtain feature differences between multiple channels. This method constructs an Infrared Differential Analysis Factor (IDAF) model to represent target-background contrast and introduces a Dual-Band Signal-to-Noise Ratio (DBSCR) to assess target observability. Finally, the IDAF-DBSCR objective function is used to select the optimal dual-band, enhancing space-based infrared target detection performance.

Liu et al. [35] proposed a fusion recognition method based on CNN to improve the recognition rate of single-band images of ships in complex backgrounds. This method integrates image data from three bands: visible light, mid-wave infrared, and long-wave infrared, to extract feature information of ship targets. Experimental results show that this method achieves a recognition accuracy of up to 84.5%, showing a significant improvement compared to single-band recognition methods. Li et al. [36] used RetinaNet, Cascade R-CNN, and CenterNet to recognize multi-band infrared ship images. The results show that the Cascade R-CNN approach demonstrated high recognition accuracy, RetinaNet excelled in achieving high precision and fast processing speed, while CenterNet showed faster running speed. Zhang et al. [37] proposed an improved deep learning object detection algorithm based on YOLOv7. By introducing attention mechanisms and optimizing the loss function, they developed a new framework that utilizes enhanced D-S evidence theory to fuse visible light, infrared, and SAR images. This framework achieves more accurate and efficient detection of camouflaged targets while meeting real-time requirements. Yin et al. [38] proposed an adaptive visual enhancement infrared-visible image fusion scheme based on high-saliency target detection, effectively addressing the issue of poor fusion between visible light and infrared images under low illumination conditions. Dahai et al. [39] proposed a decision-level fusion target detection approach for visible light and infrared images based on model reliability, which reduces the target miss detection rate during the day and night compared to single target detection models.

Multiband target detection can provide richer information than single-band detection in certain situations, but effectively fusing data from different bands remains a major challenge. Specifically, for detecting small and dim targets in infrared images, the low signal-to-noise ratio and the strong infrared radiation and complex texture structures in the background pose significant difficulties for accurate target detection. This paper integrates features from long-wave and medium-wave infrared images and proposes an end-to-end detection framework combining attention mechanisms with a self-distillation mechanism, which offers a potential approach to address this problem.

## 2.2. Attention Mechanism

With the development of attention mechanisms, their application in infrared target detection has deepened. By designing and optimizing attention mechanisms, infrared target detection approaches can significantly improve detection accuracy, enhance the ability to detect small and dim targets, and optimize computational resources. Zheng et al. [40] developed a multi-stage image fusion network (MSFAM) based on attention mechanisms for feature fusion of infrared and visible light images. This network enhances image features through attention fusion units, uses a semantic constraint module and a pull-push loss function for fusion tasks, and achieves better performance across various evaluation metrics on public datasets. Fu et al. [41] proposed a feature-enhanced remote attention fusion network (LRAF-Net), which fuses the long-range dependence of the enhanced visible and infrared features. This network has achieved more advanced detection performance compared to other methods on multiple public datasets, including VEDAI, FLIR, and LLVIP. Zhang et al. [42] introduced attention mechanisms into spatial infrared target recognition, proposing a new spatial infrared target classification framework (MCA-CNN) based on a multi-channel CNN. This network integrates multiple single-band features and uses channel attention mechanisms to assess the importance of features from different channels. In experiments, it effectively handles issues such as noise interference and missing data. Li et al. [43] proposed a fusion method based on cross-attention mechanisms (CAM) called CrossFuse, which enhances internal features of each modality with self-attention while using a cross-attention-based architecture to strengthen complementary information between different modalities. By incorporating CAM into the transformer architecture, their method provides a powerful fusion strategy for multimodal images, effectively enhancing the dissimilarity between inputs. Xiaofeng et al. [44] proposed an infrared target detection method based on a parallel attention mechanism for ground background targets. This method uses convolution and attention in parallel for down-sampling, and during multi-scale feature fusion, it reuses and complements information from multiple different scales. In comparison experiments with YOLOv3, the method improves target detection accuracy while reducing model complexity and training time. Wang et al. [45] proposed an attention-based infrared small target detection method that utilizes both passive and active attention information. The active attention module performs complementary fusion of multi-dimensional feature information through a top-down approach, while the passive attention module locates candidate target regions using a bottom-up approach with a fast local contrast algorithm and circular pipeline filtering. Subsequently, a decision-feedback balancing mechanism is employed to identify the actual targets. IR-TransDet, proposed by Lin et al. [46], integrates the advantages of Convolutional Neural Network (CNN) and Transformer, enabling precise extraction of infrared small target features and demonstrating strong performance across multiple datasets. To address the potential loss of important target and background information in the original image as the network deepens, Zhan et al. [47] proposed a Dilated Convolutional Channel Attention (DCA) module, enabling the network to better fuse contextual semantics while capturing abstract texture details and predicting distortions.

Currently, the application of attention mechanisms in the field of infrared small and dim target detection is increasing, providing solutions for infrared dual-band target detection tasks. However, extracting features from blurred and less apparent small targets amidst complex and variable background interference and effectively fusing features from two different modalities to ensure that the attention mechanism can comprehensively capture contextual information from both remains a challenge. Additionally, attention mechanisms often require significant computational resources and time for task processing. The attention module designed in this paper introduces dynamic weights for adaptive feature fusion, uses convolution operations to reduce algorithmic complexity, and extends the model's global perception capabilities.

### 3. Methods

#### 3.1. Overall Architecture

The MM-IRSTD method proposed in this research is shown in Figure 1, and its core architecture consists of two main components: the encoder and the decoder. At the input end, infrared medium-wave and long-wave infrared images are fed into a pre-trained backbone network for feature extraction, obtaining deep encoding features. Each encoder contains 5 output scales of different dimensions. For the input image, the five output feature maps are  $16 \times H/2 \times W/2$ ,  $24 \times H/4 \times W/4$ ,  $32 \times H/6 \times W/6$ ,  $96 \times H/8 \times W/8$ , and  $320 \times H/16 \times W/16$  (channels  $\times$  height  $\times$  width), where H and W represent the height and width of the image. During training, the KL divergence-based self-distillation mechanism proposed in this paper imposes spatial and dimensional constraints on the features extracted by the encoder, thereby reducing the model size and computational overhead while maintaining high accuracy. For different encoder feature maps, the self-attention module proposed in this paper performs adaptive feature fusion of the extracted long-wave and medium-wave infrared image features, which are then processed by the decoder for feature recovery and target prediction. Additionally, during model training, we embed the Sobel operator to construct boundary loss for optimizing the spatial boundary information of the saliency map.

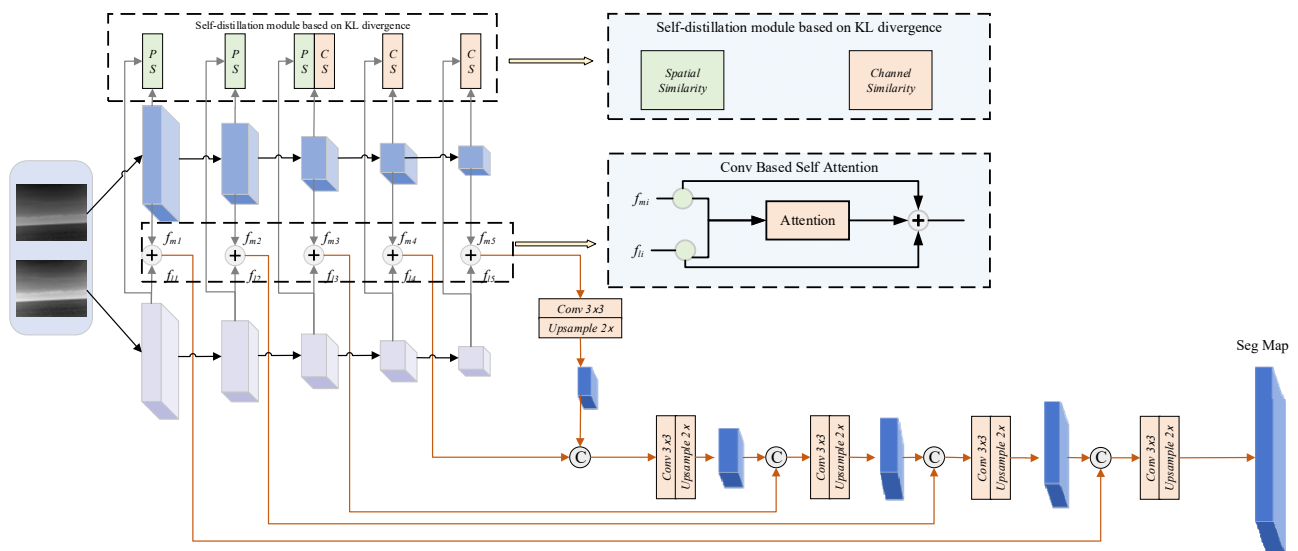


Figure 1. MM-IRSTD structure.

#### 3.2. Conv-Based Self-Attention Module

The self-attention mechanism of Transformers endows the model with global perceptual capabilities and has been widely applied in various visual tasks. In this research, we propose a Convolution-Based Self-Attention Module, which integrates channel attention mechanisms and multi-layer perceptrons. Through residual connections, this module ensures that input features are preserved and enhanced at each layer, which helps capture complex feature representations and improve model performance. As shown in Figure 2, the attention mechanism is applied to the medium-wave and long-wave image features, and the results are added to the medium-wave image features before being fed into the multi-layer perceptron. The results are then added to the input of the multi-layer perceptron to obtain the final output. In the attention module, the shape of the input Q (Query) (batch size B, number of channels C, height H, and width W) is used to generate Q via a  $1 \times 1$  convolution. Similarly, K (Key) and V (Value) are also generated using a  $1 \times 1$  convolution and are split into two parts along the channel dimension. The Q, K, and V are then reshaped to fit the multi-head attention mechanism. The Q is scaled to prevent large values, the dot product between Q and K is computed, and the result is normalized using *softmax*.

The attention-weighted  $V$  is then computed, reshaped to the original input shape, and projected through a convolution layer to produce the final output. The multi-layer perceptron module includes several convolutional layers, corresponding BatchNorm, and GELU activation functions, which perform complex linear and nonlinear transformations on the input features to extract and fuse multi-scale feature information, thereby enhancing the model's expressiveness and performance. The combination of the multi-layer perceptron with the attention mechanism further processes and strengthens the features output by the attention mechanism, enabling the model to better understand and handle the input data.

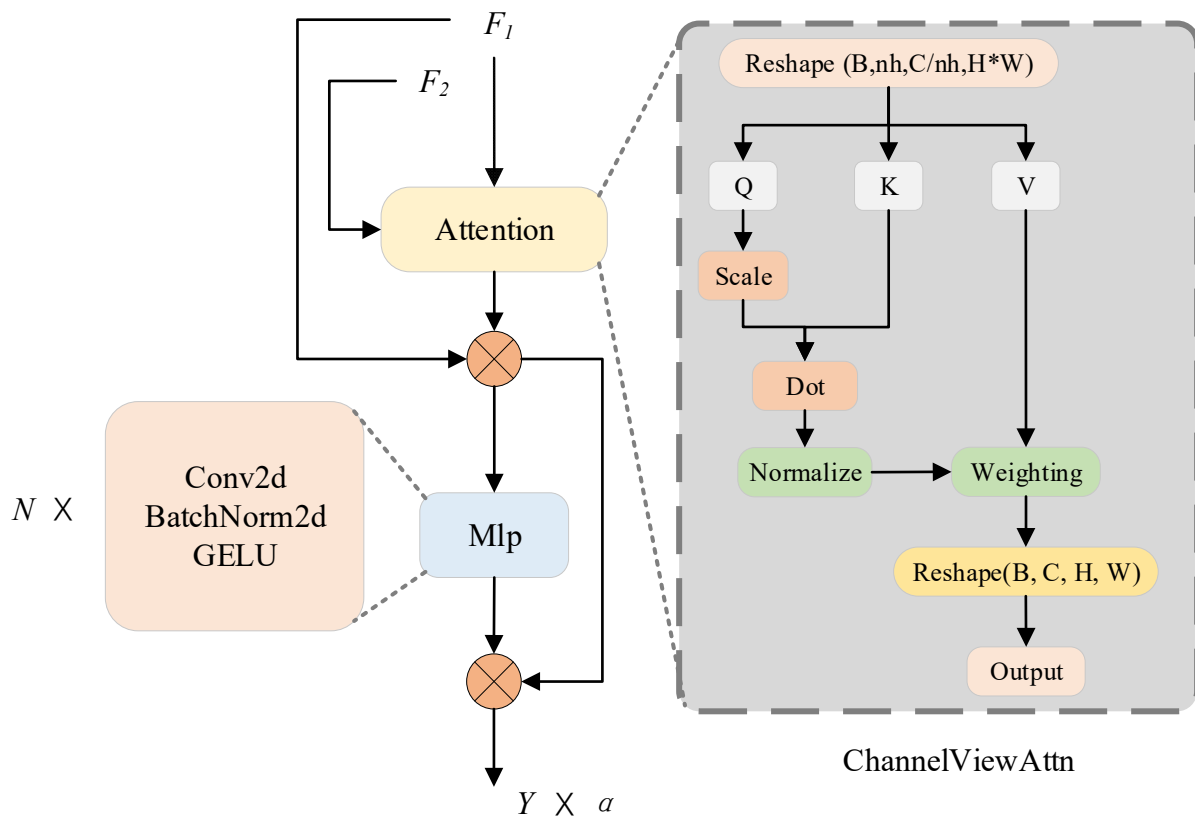


**Figure 2.** Conv-based self-attention module schematic.

The above process can be summarized by the following formula:

$$\begin{aligned}
 X &= \text{Attention}(F_1, F_2) \\
 X_1 &= F_1 + X \\
 X_2 &= \text{MLP}(X_1) \\
 Y &= (X_1 + X_2) * \alpha
 \end{aligned} \tag{1}$$

where  $F_1, F_2$  are the features of the medium-wave and long-wave input images; Attention refers to the attention mechanism;  $X_1, X_2$  are intermediate variables;  $\alpha$  represents the learnable dynamic weights, which are adaptively optimized during model training to enhance the fusion capability of multimodal features; MLP stands for multi-layer perceptron; and  $Y$  is the output.

### 3.3. Boundary Extraction Method Based on Sobel

Boundary information of targets in infrared images plays a crucial role in detecting small and dim targets. By extracting boundary information, feature representation can be improved and detection accuracy enhanced. Given that the Sobel operator is simple to compute, performs well in detection, and has good noise resistance, this paper designs a boundary extraction method based on Sobel.

As shown in Figure 3, the Sobel operator uses two  $3 \times 3$  convolution kernels,  $k_x$  and  $k_y$ , to compute the gradients of the image grayscale values in the horizontal and vertical directions, respectively. The absolute values of the horizontal and vertical gradients are then summed to obtain the edge-strength map of the image. After extracting edge information using the Sobel operator, the edge information of the target can be compared, and edge loss can be calculated to guide the model in better learning and detecting edge features in the image. This can be summarized by the following formula:

$$\begin{aligned} sobelx &= conv2d(In, weightx) \\ sobely &= conv2d(In, weighty) \\ out &= abs(sobelx) + abs(sobely) \end{aligned} \quad (2)$$

where  $In$  is the input image tensor,  $weightx$  and  $weighty$  are the convolution kernels in the horizontal and vertical directions, respectively,  $conv2d$  denotes the convolution operation,  $abs$  is the absolute value operation, and  $out$  is the overall edge output.

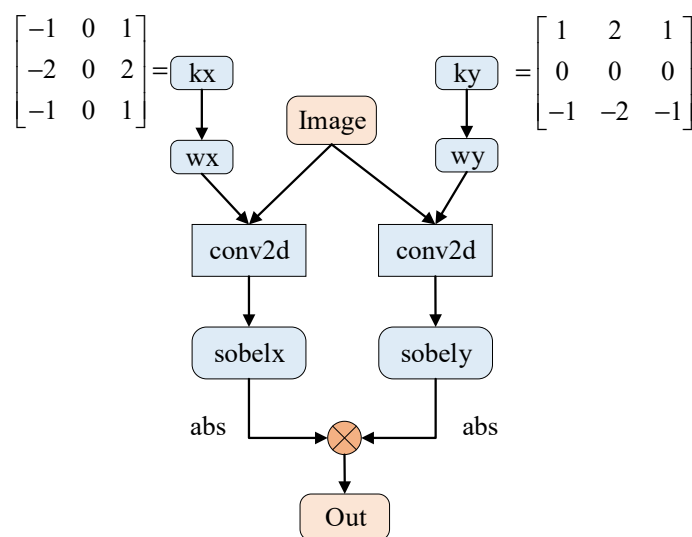


**Figure 3.** Schematic of the Sobel operator.

### 3.4. KL Divergence-Based Self-Distillation Mechanism

Knowledge distillation, as a key technology for model optimization and performance enhancement, focuses on effectively transferring the core knowledge from a large and complex model (the teacher model) to a more compact model (the student model). This process aims to enable the student model to exhibit performance comparable to that of the teacher model while significantly reducing computational resource consumption. The teacher model is usually a complex, large model that requires significant computational resources and time for training. In this paper, we design a self-distillation mechanism combining KL divergence and similarity constraints. This module is active only during model training and can be automatically discarded during inference, thereby ensuring better inference performance. The module includes two components: spatial similarity constraint and channel similarity constraint.

#### 3.4.1. Spatial Similarity Constraints

Significant progress has been made in infrared small and dim target detection using various machine learning strategies. Although many methods have enhanced performance through complex model designs, this often results in a substantial increase in the number of parameters. Our goal is to leverage multimodal information while keeping the model size unchanged, thereby reducing test time and ensuring efficiency.

In infrared weak small target detection methods, low-level feature representation spaces and structural information are crucial. Therefore, we propose a Spatial Similarity

Constraint Module (PS) to leverage attention transfer learning of low-level spatial features from long-wave and mid-wave images.

We use the following definitions and symbols. Features in the long-wave image are denoted as  $l_i$ , while features from the mid-wave image are denoted as  $m_i$ . The feature map dimensions are given by the number of channels  $\times$  height  $\times$  width, where  $i = 1, 2, 3$ , and  $l_i$  and  $m_i$  have the same structure. We illustrate the process of attentive feature transfer learning using the example of transferring features from  $l_i$  to  $m_i$ . Initially, we implement spatial attention transfer learning through a convolution and the sigmoid function. Subsequently, during the attentive feature transfer process, global normalization is employed to identify and differentiate the varying levels of importance of different regions in the feature map on the two-dimensional plane. Finally, the process from  $l_i$  to the spatial attention weights  $w_{li}$  can be represented as follows:

$$w_{li} = \text{globnorm}(\zeta(\text{Conv}(l_i^{\text{detach}}))), i = 1, 2, 3 \quad (3)$$

where  $l_i^{\text{detach}}$  is the gradient that removes  $l_i$  while retaining its value, allowing  $m_i$  to learn  $l_i$  in a specific direction; Conv represents the convolution operation;  $\zeta$  is the sigmoid function; and  $\text{globnorm}$  is the global normalization operation.

To ensure that the features are within a similar range and to facilitate the learning process, we follow the approach in [48] and apply the L2 norm to  $l_i^{\text{detach}}$  and  $m_i$  before transfer learning:

$$\begin{aligned} l_{i\_norm}^{\text{detach}} &= \text{L2norm}(l_i^{\text{detach}}), i = 1, 2, 3 \\ m_{i\_norm} &= \text{L2norm}(m_i), i = 1, 2, 3 \end{aligned} \quad (4)$$

where  $\text{L2norm}$  is the L2 norm operation, which is performed along the channel dimension to preserve the geometric information of the spatial structure.

The loss from  $l_i$  to  $m_i$  is represented by the sum of mean squared error and KL divergence:

$$\ell_{l \rightarrow m}^{\text{PS}} = \sum_{i=1}^3 w_{li} \times \text{MSE}(m_{i\_norm}, l_{i\_norm}^{\text{detach}}) + \text{KLdiv}(m_{i\_norm}, l_{i\_norm}^{\text{detach}}), i = 1, 2, 3 \quad (5)$$

where  $\text{MSE}$  denotes the mean squared error loss, and  $\text{KLdiv}$  is used to compute the KL divergence-based loss.

As shown in Figure 4, the geometric transfer loss calculation process from  $l_i$  to  $m_i$ .

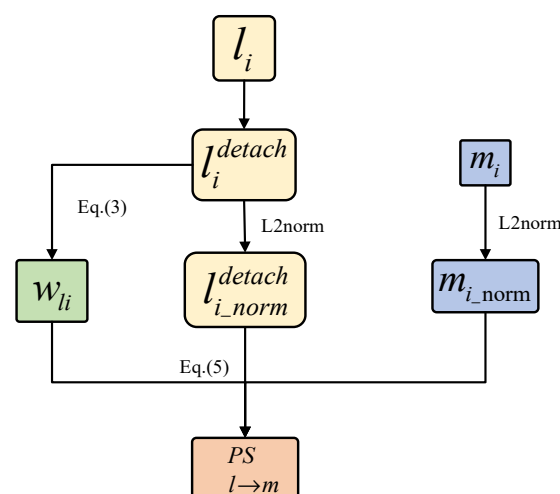


**Figure 4.** Geometric transfer loss from  $l_i$  to  $m_i$ .

The PS from  $m_i$  to  $l_i$  is the same as the PS from  $l_i$  to  $m_i$ ; therefore, we directly express the PS from  $m_i$  to  $l_i$  as follows:

$$\ell_{m \rightarrow l}^{PS} = \sum_{i=1}^3 w_{mi} \times \text{MSE}(l_{i\_norm}, m_{i\_norm}^{detach}) + \text{KLdiv}(l_{i\_norm}, m_{i\_norm}^{detach}), i = 1, 2, 3 \quad (6)$$

The sum of  $\ell_{m \rightarrow l}^{PS}$  and  $\ell_{l \rightarrow m}^{PS}$  is the final geometric transfer loss, denoted as  $\ell^{PS}$ .

$$\ell^{PS} = \ell_{m \rightarrow l}^{PS} + \ell_{l \rightarrow m}^{PS} \quad (7)$$

### 3.4.2. Channel Similarity Constraints

Both CS (Channel similarity constraints) and PS involve transfer learning, with CS emphasizing high-level semantic information. The main difference between CS and PS lies in the computation of weights and dimensions using the L2 norm.

Taking the example of transferring features from  $l_i$  to  $m_i$ , the feature map dimensions are also channel  $\times$  height  $\times$  width. For CS, we use channel attention to replace the spatial attention in PS, and use two convolutions (Conv), global average pooling (GAP), and a sigmoid function ( $\zeta$ ) to generate a global semantic weight. In particular, global normalization (*globnorm*) first reduces the feature map to a vector of channel dimensions through GAP and then normalizes along this dimension to aggregate and standardize the semantic information of each channel. The L2 norm operation in CS is performed on the planar level to preserve the semantic representation of individual channels.

Similar to Equation (7), the semantic transfer loss ( $\ell^{CS}$ ) is given by:

$$\ell^{CS} = \ell_{m \rightarrow l}^{CS} + \ell_{l \rightarrow m}^{CS} \quad (8)$$

where  $\ell_{m \rightarrow l}^{CS}$  represents the CS from  $m_i$  to  $l_i$ , and  $\ell_{l \rightarrow m}^{CS}$  represents the CS from  $l_i$  to  $m_i$ , both of which are the sum of mean squared error and KL divergence.

### 3.4.3. Invocation of the Module

The use of the MM-IRSTD with self-distillation structure is shown in Algorithm 1. After initializing the model, MM-IRSTD extracts features from long-wave and mid-wave images, then performs feature fusion through a multi-head attention mechanism. Next, it upsamples and concatenates the feature maps, gradually upsampling until the final output features are obtained. In training mode, the self-distillation mechanism calculates the loss under spatial similarity constraints and channel similarity constraints, returning the loss value along with the final output. In addition, the self-distillation loss in the training mode is summed with losses such as IOU, to drive model updates. If not in training mode, the self-distillation structure will not be called, and only the final output will be returned.

---

#### Algorithm 1. Invocation process of the self-distillation mechanism

---

Step 1: Extract long-wave image features and medium-wave image features

Step 2: Fuse features

Step 3: Upsample and feature concatenation

if training

calculate attention weight  $w_{li} = \text{globnorm}(\zeta(\text{Conv}(I_i^{detach}))), i = 1, 2, 3$

apply L2 norm, get

$l_{i\_norm}^{detach} = L2norm(I_i^{detach}), i = 1, 2, 3$

$m_{i\_norm} = L2norm(m_i), i = 1, 2, 3$

use MAE and KL divergence to get  $\ell_{l \rightarrow m}^{PS}$

same way, get  $\ell_{m \rightarrow l}^{PS}$ , then  $\ell^{PS} = \ell_{m \rightarrow l}^{PS} + \ell_{l \rightarrow m}^{PS}$

calculate  $\ell^{CS} = \ell_{m \rightarrow l}^{CS} + \ell_{l \rightarrow m}^{CS}$

return  $\ell^{PS} + \ell^{CS}$ , out

else

Not invoking the self-distillation mechanism to improve inference speed

return out

---

### 3.5. Loss Function

The loss functions used in model training include IOUBCE and IoU loss. The IOUBCE loss function combines binary cross-entropy (BCE) loss with intersection-over-union (IoU) loss. BCE loss is used to compute classification errors, ensuring that the probability output by the model is as close as possible to the true labels. IoU loss is used to calculate the overlap between the predicted segmentation region and the true segmentation region, optimizing the accuracy of the segmentation boundaries. This approach considers both pixel-level errors and the overlap of target regions.

The BCE calculation formula is as follows:

$$\text{BCE}(P, T) = -\frac{1}{N} \sum_{i=1}^N [T_i \log(\sigma(P_i)) + (1 - T_i) \log(1 - \sigma(P_i))] \quad (9)$$

where  $N$  is the total number of pixels,  $P$  is the predicted value,  $T$  is the true value, and  $\sigma$  is the Sigmoid activation function.

The IoU loss is calculated as:

$$\text{IoU}(P, T) = \frac{\sum_{i=1}^N [\sigma(P_i) \cdot T_i] + 1}{\sum_{i=1}^N [\sigma(P_i) + T_i] - \sum_{i=1}^N [\sigma(P_i) \cdot T_i] + 1} \quad (10)$$

Then, the final IOUBCE loss function is

$$\text{IOUBCE}(P, T) = \frac{1}{B} \sum_{j=1}^B (1 - \text{IoU}(P, T) + \text{BCE}(P, T)) \quad (11)$$

where  $B$  is the input sample size.

The IoU loss is calculated as:

$$\text{IoU Loss}(P, T) = 1 - \frac{1}{N} \sum_{i=1}^N \frac{\sum [\sigma(P_i) \cdot T_i] + \varepsilon}{\sum [\sigma(P_i) + T_i] - \sum_{i=1}^N [\sigma(P_i) \cdot T_i] + \varepsilon} \quad (12)$$

where  $\varepsilon$  is the smoothing term, which usually takes the value 1.

## 4. Experimental

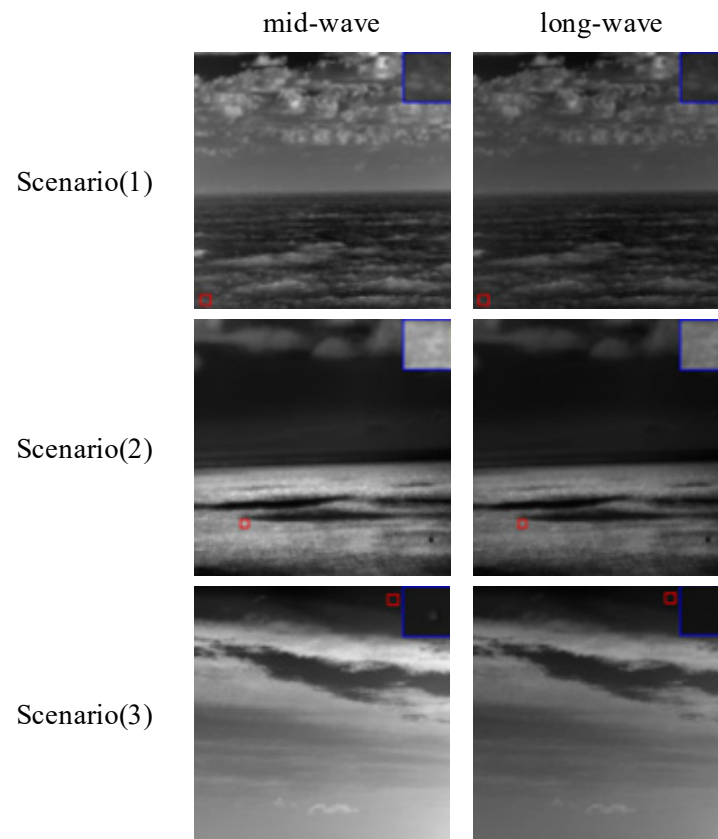
### 4.1. Dataset

The dataset used for model validation in this paper is sourced from a self-constructed dataset by the research group, composed of real background and simulated targets. The target type is point targets, and each image contains a target. The data cover various complex background interferences, including high-light sea clutter and complex cloud backgrounds. The dataset includes 15,331 pairs of strictly aligned medium-wave and long-wave images, with 12,214 pairs used for training and 3117 pairs used for testing, as shown in Table 1. The long-wave images are captured using a  $512 \times 512$  specification long-wave infrared thermal camera, while the medium-wave images are captured using a  $512 \times 512$  specification mid-wave infrared thermal camera, with a frame rate of 50 Hz. The long-wave wavelength is 8–12  $\mu\text{m}$ , and the mid-wave wavelength is 3–5  $\mu\text{m}$ . The training and test sets do not come from the same sequence and are completely non-overlapping, resulting in differing backgrounds. Additionally, the test set has a higher signal-to-noise ratio, making detection more challenging. We use a more rigorous test set to evaluate the performance of the designed network.

As shown in Figure 5, here are some images from the dataset, where the red box contains the target, and the blue box shows the enlarged view of the target.

**Table 1.** Dataset details.

|                     | Size of Images | Number of Images | Types of Background                       | Size of the Target | Signal-to-Noise Ratio      |
|---------------------|----------------|------------------|-------------------------------------------|--------------------|----------------------------|
| <b>Training Set</b> | 512 × 512      | 12,214           | Clouds/city/sea/field/river/<br>mountains | 10–15              | 0.399625(l)<br>1.612668(m) |
| <b>Test Set</b>     | 512 × 512      | 3117             | Clouds/sea/field                          | 10–15              | 0.305886(l)<br>0.640464(m) |

**Figure 5.** Sample of the dataset. The red box contains the target, and the blue box shows the enlarged view of the target.

#### 4.2. Metrics

This paper uses several evaluation metrics to assess the detector's performance, including IoU, Dice coefficient, Precision, Recall, F1-score, and Kappa statistic. IoU measures the overlap between the ground truth and predicted bounding boxes, calculated as the intersection divided by the union. The Dice coefficient evaluates set similarity, reacting to the degree of similarity between two sets. Precision indicates the proportion of true positives among all predicted positives, while Recall measures the proportion of true positives identified out of all actual positives. The F1-score provides a balanced measure by combining Precision and Recall. Finally, Kappa assesses the agreement between model predictions and ground truth, adjusting for chance agreements. The detailed mathematical formulas for these metrics are provided below:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (13)$$

$$\text{Dice} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (14)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (15)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (16)$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

where TP represents the number of pixels correctly identified as the target; TN represents the number of pixels correctly identified as the background; FP denotes the number of pixels incorrectly identified as the target instead of the background; and FN represents the number of pixels incorrectly identified as the background instead of the target.

$$\text{kappa} = \frac{p_o - p_e}{1 - p_e} \quad (18)$$

$$\text{where } p_o = \frac{TP+TN}{TP+TN+FP+FN}, p_e = \frac{(TP+FN) \cdot (TP+FP) + (FP+TN) \cdot (FN+TN)}{(TP+TN+FP+FN)^2}.$$

#### 4.3. Implementation Details

The parameter configuration for the MM-IRSTD approach is as follows: the initial learning rate is 0.0001, the batch size is 16, the epochs is 20, and the optimizer used is Adam. For image feature extraction, MobileNetV2 is chosen as the backbone network. In the Path Transformer, the number of blocks is set to 4–8. The computing device is a single Nvidia RTX 3090, and the PyTorch version is 2.0.

#### 4.4. Model Training

We measured the loss function values of both the training model and the testing model after each round of network training. Figure 6 shows the variation of the loss function values during model training and testing with respect to epochs. It can be observed that as the number of epochs increases, both the training loss and the testing loss gradually approach 0, indicating that the MM-IRSTD model training and testing are converging. Moreover, the downward trends of the training loss and testing loss are consistent, indicating that there is no overfitting issue.

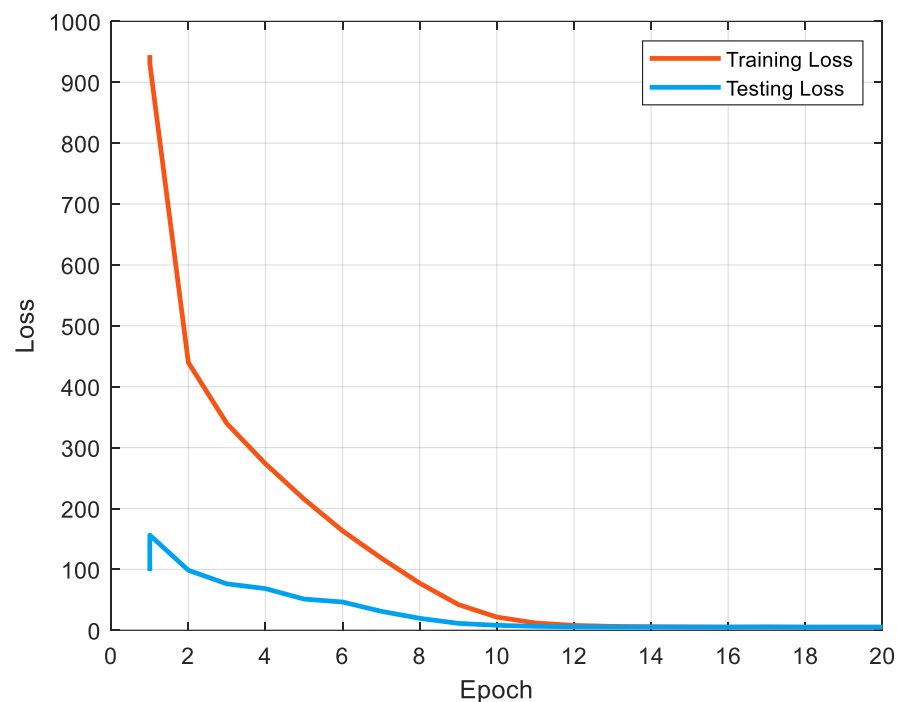


**Figure 6.** The variation of training loss and testing loss with epochs.

#### 4.5. Comparison Experiment

To test the performance of the MM-IRSTD on dual-band infrared small and dim target detection tasks, this paper selects six comparative methods: four state-of-the-art (SOTA) RGB-T salient object detection approaches, CAVER50 (with ResNet50D), CAVER101 (with ResNet101D), LSNet, and SPNet, and two infrared small and dim target detection methods based on long-wave images, ACM and DNANet, which use a gray background for highlighting in Table 2.

**Table 2.** SOTA method comparison results. The best-performing method is highlighted in red, while the second-best method is highlighted in blue.

| Model       | IoU          | Dice         | Precision    | Recall | Fscore       | Kappa        | Fps          |
|-------------|--------------|--------------|--------------|--------|--------------|--------------|--------------|
| ACM         | 0.711        | 0.797        | 0.877        | 0.745  | 0.797        | 0.594        | 25.18        |
| DNANet      | 0.688        | 0.774        | 0.900        | 0.708  | 0.774        | 0.548        | 4.43         |
| CAVER 50    | 0.734        | 0.819        | 0.776        | 0.877  | 0.819        | 0.638        | 22.49        |
| CAVER 101   | 0.746        | 0.830        | 0.787        | 0.887  | 0.830        | 0.660        | 18.61        |
| LSNet       | 0.745        | 0.829        | 0.783        | 0.892  | 0.829        | 0.658        | 43.01        |
| SPNet       | 0.670        | 0.754        | 0.676        | 0.954  | 0.754        | 0.508        | 20.01        |
| <b>Ours</b> | <b>0.776</b> | <b>0.856</b> | <b>0.929</b> | 0.804  | <b>0.856</b> | <b>0.711</b> | <b>45.84</b> |

##### 4.5.1. Quantitative Analysis

Table 2 presents the test results of various methods, where the methods with a gray background use dual-band images, and those with a white background, including our method, use dual-band images. The results show that MM-IRSTD achieves better performance. In terms of IoU and Dice metrics, MM-IRSTD outperforms other methods, indicating that the detection maps produced by MM-IRSTD have a higher correlation with the ground truth maps. MM-IRSTD enhances the accuracy of predicted image shapes by cascading multiple FFU modules and repeatedly fusing multi-band features, significantly improving the precise recognition of small and dim target shapes, which can notably enhance the accuracy of subsequent tasks, such as improving the targeting precision of early warning systems and the ability to perceive target scales. Precision and Recall, respectively, represent the method's positive predictive value and the completeness of detection.

Compared to dual-band image detectors, ACM and DNANet using medium-wave images show better results in Precision; however, these methods significantly lag in Recall compared to other methods. This is mainly because single-band detectors, limited to using only one band of image data, cannot effectively integrate multi-band information to mitigate background noise interference, resulting in lower recall rates, where fewer targets are recognized and detected, leading to more missed detections. In terms of overall evaluation metrics, F-score and Kappa, MM-IRSTD improves by 3.13% and 7.73%, respectively, compared to the second-ranking method, CAVER 101 (with ResNet101D as the backbone). In terms of method speed, MM-IRSTD performs the best, outperforming the second-ranked method, LSNet, by 6.58%, demonstrating its excellent inference speed.

##### 4.5.2. Qualitative Evaluation

We selected six SOTA approaches for qualitative analysis, including two detection approaches that process only medium-wave images (ACM and DNANet), and four additional image analysis approaches designed for complex scenes (GT, CAVER, LSNet, SPNet). By applying these approaches to the actual collected dataset and comparing their performance, we achieved the results shown in Figure 7.

Experimental analysis shows that MM-IRSTD demonstrates outstanding target recognition capability across various scenes. It not only achieves more precise edge detection but also significantly reduces false detection rates, while attaining higher accuracy in target localization. Among the comparison methods, ACM and DNANet exhibit varying degrees of erroneous detection in environments with complex fish-scale glare interference. Due to the prominent fish-scale glare interference in medium-wave images, these methods, which

rely solely on single-band image information, struggle to suppress complex and variable background interference, leading to difficulties in accurately and effectively recognizing targets in complex scenes. Among multi-band detection methods, LSNet performs poorly, primarily due to its use of a simple pixel stacking method for handling multi-band features. This method limits the method's ability to delve into higher-order information and effectively suppress background interference, resulting in noticeable false detections and missed detections. In contrast, the SPNet cleverly leverages the saliency features of images to enhance semantic information. In the experimental results, this method achieved the best performance in Recall. However, SPNet does not adequately consider the importance of feature differences between different modalities in its design. The model's overemphasis on feature extraction leads to noticeable false detection issues during the detection process.

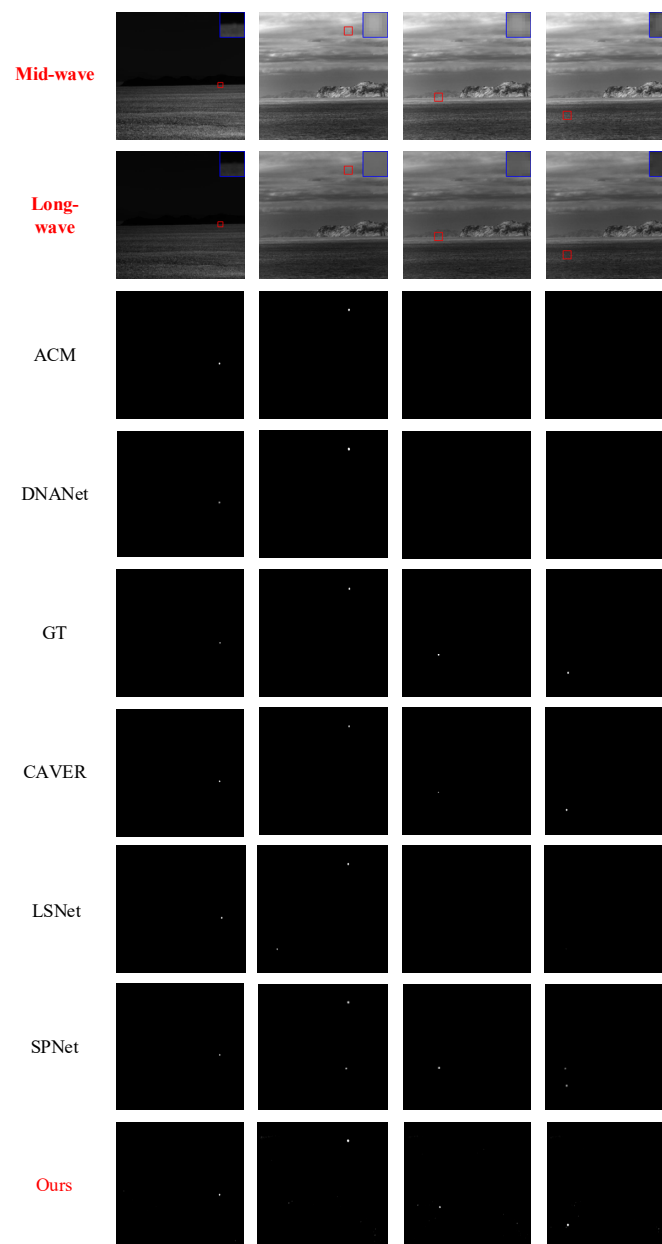


**Figure 7.** Qualitative analysis of results. Mid-wave represents the mid-wave image, while Long-wave represents the long-wave image. The red box indicates the target, and the blue box shows the situation after the target has been magnified.

#### 4.6. Ablation Experiments

This paper conducts ablation experiments on the MM-IRSTD framework using a self-constructed dataset from the laboratory. Specifically, we investigated how different backbone networks, as well as the proposed Conv-Based Self-Attention module and KL-divergence-based self-distillation mechanism, affect the performance of the MM-IRSTD.

##### 4.6.1. Model Framework Ablation Experiment

We tested the contribution of different modules to the final detection results by shielding various parts of the MM-IRSTD. The experimental results are shown in Table 3, with the best performance is highlighted in red. Exp4 represents the complete method framework. In Exp1, the Sobel operator was shielded. The Sobel module in this framework computes the gradients of the image in the x and y directions and returns the gradient magnitude map. Compared to Exp4, Exp1 showed a 4.38% and 3.63% decrease in IoU and F1-score, respectively, making it the module with the largest decrease when shielded, demonstrating the importance of the Sobel module in the MM-IRSTD framework. In Exp2, the Conv Based Self Attention module was shielded. This module enhances the representation ability of input features through a combination of self-attention mechanisms and multi-layer perceptrons, while residual connections improve the model's training stability and performance.

**Table 3.** Analysis of the validity of the modules proposed in the MM-IRSTD. The maximum value of the metric is highlighted in red.

| Exp | Model                  | IoU   | Dice  | Precision | Recall | F-Score | Kappa |
|-----|------------------------|-------|-------|-----------|--------|---------|-------|
| 1   | No Sobel               | 0.742 | 0.826 | 0.873     | 0.789  | 0.826   | 0.652 |
| 2   | No Transformer         | 0.745 | 0.829 | 0.783     | 0.892  | 0.829   | 0.658 |
| 3   | No KL                  | 0.770 | 0.851 | 0.860     | 0.849  | 0.851   | 0.709 |
| 4   | Ours (Sobel + KL + TF) | 0.776 | 0.856 | 0.929     | 0.804  | 0.856   | 0.711 |

Compared to Exp4, Exp2 showed a 3.99% and 3.15% decrease in IoU and F1-score, respectively, with a decrease similar to Exp1, further confirming the significant contribution of this module to the MM-IRSTD framework. Exp3 shielded the KL-divergence-based self-distillation mechanism, which was used only during the training process. This module facilitates the transfer from the teacher model to the student model by constraining the spatial and channel features of the feature maps. Compared to Exp4, Exp3 showed the smallest decrease of 0.77% and 0.58% in IoU and F1-score, respectively, confirming the module's effectiveness in enhancing the performance of the framework for detecting small and dim targets.

##### 4.6.2. Model Parameter Ablation Experiment

In MM-IRSTD, the number of input channels is gradually reduced from 320 to 16 to progressively downsample the high-level feature maps, while the number of attention heads is configured as (4, 4, 4, 6, 8). Generally, the number of channels should be divisible by the number of attention heads to ensure that each head can process the input features evenly. Too few heads may fail to fully leverage the advantages of multi-head attention, while too many can lead to unnecessary computational overhead. To demonstrate the rationality of the head count configuration used in this paper, we tested the configurations (4, 4, 4, 4, 4), (4, 4, 8, 8, 8), (4, 4, 4, 8, 8), and (4, 8, 8, 8, 8), with the comparative results shown in Table 4. The results indicate that Exp5 performs the best in all metrics except Precision. In terms of IoU and F-score, Exp5 shows improvements of 5.58% and 4.39%, respectively, compared to the second-best Exp1. Thus, we recognize that different scales correspond to different numbers of self-attention heads, enhancing the model's multi-scale perception capability.

**Table 4.** Results of the model parameter ablation study. The maximum value of the metric is highlighted in red.

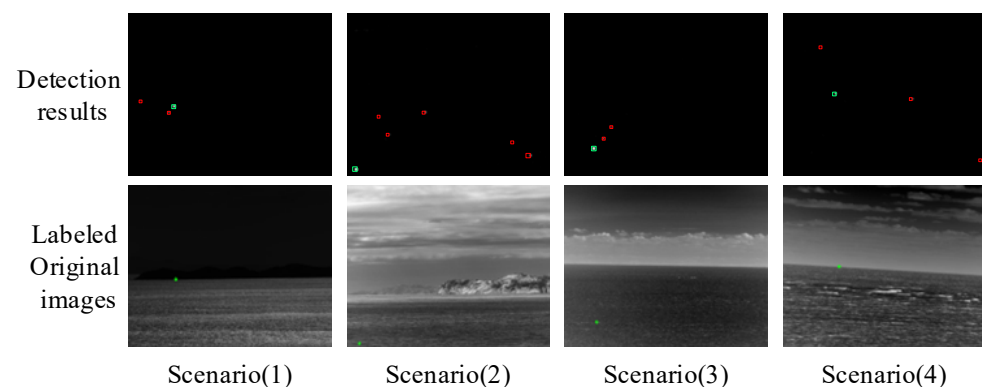
| Exp | Parameters       | IoU   | Dice  | Precision | Recall | F-Score | Kappa |
|-----|------------------|-------|-------|-----------|--------|---------|-------|
| 1   | (4,4,4,4,4)      | 0.735 | 0.820 | 0.937     | 0.752  | 0.820   | 0.639 |
| 2   | (4,4,8,8,8)      | 0.732 | 0.817 | 0.953     | 0.744  | 0.817   | 0.634 |
| 3   | (4,4,4,8,8)      | 0.714 | 0.800 | 0.938     | 0.728  | 0.800   | 0.599 |
| 4   | (4,8,8,8,8)      | 0.728 | 0.813 | 0.890     | 0.762  | 0.813   | 0.627 |
| 5   | Ours (4,4,4,6,8) | 0.776 | 0.856 | 0.929     | 0.804  | 0.856   | 0.711 |

## 5. Discussion

In this paper, we introduce an attention mechanism to perform feature fusion between long-wave infrared and mid-wave infrared images, use the Sobel operator to enhance the model's edge detection capability, and build a KL-divergence-based self-distillation mechanism to achieve model lightweighting. As shown in Table 2 and Figure 7, the MM-IRSTD outperforms state-of-the-art (SOTA) methods in both quantitative and qualitative comparison experiments.

Through ablation experiments on MM-IRSTD, we find that the Sobel operator has the greatest contribution to detection results, followed by the attention mechanism, while the KL-divergence-based self-distillation mechanism has the least contribution. This is because the Sobel operator significantly improves the overall model's performance and stability by enhancing edge features, reducing background interference, and simplifying computational complexity. In the fusion target detection of long-wave and mid-wave infrared images, the attention mechanism also improves the model's detection performance by enhancing feature representation, effectively fusing information, reducing redundant information, and increasing model robustness. On the other hand, the KL-divergence in the self-distillation mechanism is used merely for having the student model mimic the teacher model's output distribution and does not focus on improving detection performance, hence its lower contribution to detection results.

When conducting target detection with MM-IRSTD, we also encountered some erroneous detection results, as shown in Figure 8. This is due to the small size of the targets, while the background is filled with clouds and surface ripples. The complex background shares similar infrared characteristics with the small targets, leading to confusion between targets and background. In some images, the targets are even located in low signal-to-noise ratio regions at the horizon, which significantly increases the difficulty of target detection and results in decreased accuracy and false detections. To reduce these detection errors, further optimization of the preprocessing steps is needed to suppress the background, specifically optimize for small targets, and enhance the model's robustness.

**Figure 8.** Incorrect detection examples. In the detection results, the green box indicates correctly detected targets, while the red box indicates incorrectly detected targets. In the labeled original images, the area within the green edges represents the target.

## 6. Conclusions

This paper addresses the challenge of weak target features in the detection of small and dim infrared targets by proposing an innovative end-to-end dual-band infrared detection framework. This framework effectively integrates features from medium-wave and long-wave infrared images. To enhance feature fusion, we introduce a Conv-Based Self-Attention module that leverages attention mechanisms to weight input features, coupled with multi-layer perceptrons for deep processing and nonlinear transformation. This combination not only preserves the original feature information but also enhances the focus and representation of critical features, thereby improving the overall performance of the model. To achieve a lightweight network architecture, we propose a KL divergence-based self-distillation mechanism that constrains both spatial and channel dimensions, ensuring efficient inference. Furthermore, we incorporate Sobel kernels and convolution operations to compute gradient magnitudes and perform edge detection on the input images. Utilizing our lab's self-constructed dataset, we conduct both quantitative and qualitative comparative analyses. The proposed MM-IRSTD method demonstrates superior accuracy in the detection of small and dim infrared targets when compared to state-of-the-art (SOTA) detectors.

**Author Contributions:** Conceptualization, J.Y. and S.L.; methodology, Z.Y. and J.L.; software, J.L.; validation, J.Y., Z.Y. and J.L.; formal analysis, J.Y.; investigation, D.C.; resources, S.L.; data curation, L.D.; writing—original draft preparation, Z.Y.; writing—review and editing, Z.Y. and J.L.; visualization, Z.Y.; supervision, J.Y.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the National Natural Science Foundation of China under Grant 62273279, the National Funded Postdoctoral Researcher Program of China under Grant GZC20232105, and the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2024043.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Conflicts of Interest:** Authors Junyan Yang, Dongfang Chen and Lingbian Du were employed by the company Shanghai Academy of Spaceflight Technology. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Yang, X.; Li, S.; Niu, S.; Yan, B.; Meng, Z. Graph-based Spatio-Temporal Semantic Reasoning Model for Anti-occlusion Infrared Aerial Target Recognition. *IEEE Trans. Multimed.* **2024**, 1–15. [\[CrossRef\]](#)
2. Yang, X.; Li, S.; Zhang, L.; Yan, B.; Meng, Z. Anti-Occlusion Infrared Aerial Target Recognition with Vision-Inspired Dual-Stream Graph Network. *IEEE Trans. Geosci. Remote Sens.* **2024**, 52, 5004614. [\[CrossRef\]](#)
3. Yang, X.; Li, S.; Cai, B.; Meng, Z.; Yan, J. MF-GCN: Motion Flow-Based Graph Network Learning Dynamics for Aerial IR Target Recognition. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, 59, 6346–6359. [\[CrossRef\]](#)
4. Tian, X.; Li, S.; Yang, X.; Zhang, L.; Li, C. Joint spatio-temporal features and sea background prior for infrared dim and small target detection. *Infrared Phys. Technol.* **2023**, 130, 104612. [\[CrossRef\]](#)
5. Bai, X.; Zhou, F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognit.* **2010**, 43, 2145–2156. [\[CrossRef\]](#)
6. Chen, C.L.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. [\[CrossRef\]](#)
7. Dai, Y.; Wu, Y. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, 10, 3752–3767. [\[CrossRef\]](#)
8. Lin, J.; Li, S.; Yang, X.; Niu, S.; Yan, B.; Meng, Z. CS-ViG-UNet: Infrared small and dim target detection based on cycle shift vision graph convolution network. *Expert Syst. Appl.* **2024**, 254, 124385. [\[CrossRef\]](#)
9. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

10. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [\[CrossRef\]](#)
12. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
13. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
15. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
16. Redmon, J.; Farhadi, A. Yolo3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
17. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolo4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
18. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Michael, K.; Fang, J.; Yifu, Z.; Wong, C.; Montes, D.; et al. YOLOv5 by Ultralytics, Version 7.0. 2020. Available online: <https://github.com/ultralytics/yolov5> (accessed on 22 July 2024).
19. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YoloX: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
20. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976v1.
21. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696v1.
22. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO, Version 8.0.0. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 22 July 2024).
23. Wang, C.Y.; Yeh, I.H.; Liao, H.-Y.M. Yolo9: Learning what you want to learn using programmable gradient information. *arXiv* **2024**, arXiv:2402.13616.
24. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolo10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.
25. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part I 14. pp. 21–37.
26. Fu, C.-Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
27. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
28. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
29. Sun, Z.; Cao, S.; Yang, Y.; Kitani, K.M. Rethinking transformer-based set prediction for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3611–3620.
30. Dai, Z.; Cai, B.; Lin, Y.; Chen, J. Up-detr: Unsupervised pre-training for object detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1601–1610.
31. Zheng, M.; Gao, P.; Zhang, R.; Li, K.; Wang, X.; Li, H.; Dong, H. End-to-end object detection with adaptive clustering transformer. *arXiv* **2020**, arXiv:2011.09315.
32. Kou, T.; Zhou, Z.; Liu, H.; Yang, Y. Multi-band composite detection and recognition of aerial infrared point targets. *Infrared Phys. Technol.* **2018**, *94*, 102–109. [\[CrossRef\]](#)
33. Liu, J.; Gong, W.; Zhang, T.; Zhang, Y.; Deng, W.; Liu, H. Multi-band Image Fusion With Infrared Broad Spectrum For Low And Slow Small Target Recognition. In Proceedings of the 2022 International Conference on Artificial Intelligence and Computer Information Technology (AICIT), Yichang, China, 16–18 September 2022; pp. 1–5.
34. Guo, L.; Rao, P.; Chen, X.; Li, Y. Infrared differential detection and band selection for space-based aerial targets under complex backgrounds. *Infrared Phys. Technol.* **2024**, *138*, 105172. [\[CrossRef\]](#)
35. Liu, F.; Shen, T.; Ma, X. Convolutional neural network based multi-band ship target recognition with feature fusion. *Acta Opt. Sin.* **2017**, *37*, 1015002.
36. Li, Y.; Zhang, D.; Fan, L.; Ma, H.; Xu, Z. Performance Analysis of Ship Target Recognition in Multi-Band Infrared Images Based on Deep Learning. In Proceedings of the 2023 7th International Conference on Transportation Information and Safety (ICTIS), Xi'an, China, 4–6 August 2023; pp. 432–436.
37. Zhang, T.; Zhang, D.; Liu, Y. Research on Camouflage Target Detection Method Based on Dual Band Optics and SAR Image Fusion. In Proceedings of the International Conference on Image, Vision and Intelligent Systems, Baoding, China, 16–18 August 2023; pp. 320–335.
38. Yin, W.; He, K.; Xu, D.; Yue, Y.; Luo, Y. Adaptive low light visual enhancement and high-significant target detection for infrared and visible image fusion. *Vis. Comput.* **2023**, *39*, 6723–6742. [\[CrossRef\]](#)

39. Dahai, N.; Sheng, Z. An object detection algorithm based on decision-level fusion of visible and infrared images. *Infrared Technol.* **2023**, *45*, 282–291.
40. Zheng, X.; Yang, Q.; Si, P.; Wu, Q. A multi-stage visible and infrared image fusion network based on attention mechanism. *Sensors* **2022**, *22*, 3651. [[CrossRef](#)]
41. Fu, H.; Wang, S.; Duan, P.; Xiao, C.; Dian, R.; Li, S.; Li, Z. LRAF-Net: Long-Range Attention Fusion Network for Visible–Infrared Object Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 13232–13245. [[CrossRef](#)]
42. Zhang, S.; Rao, P.; Zhang, H.; Chen, X.; Hu, T. Spatial infrared objects discrimination based on multi-channel CNN with attention mechanism. *Infrared Phys. Technol.* **2023**, *132*, 104670. [[CrossRef](#)]
43. Li, H.; Wu, X.-J. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Inf. Fusion* **2023**, *103*, 102147. [[CrossRef](#)]
44. Zhao, X.; Xu, Y.; Wu, F.; Niu, J.; Cai, W.; Zhang, Z. Ground infrared target detection method based on a parallel attention mechanism. *Infrared Laser Eng.* **2022**, *51*, 20210290-1.
45. Wang, X.; Lu, R.; Bi, H.; Li, Y. An Infrared Small Target Detection Method Based on Attention Mechanism. *Sensors* **2023**, *23*, 8608. [[CrossRef](#)]
46. Lin, J.; Li, S.; Zhang, L.; Yang, X.; Yan, B.; Meng, Z. IR-TransDet: Infrared dim and small target detection with IR-transformer. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5004813. [[CrossRef](#)]
47. Zhan, W.; Zhang, C.; Guo, S.; Guo, J.; Shi, M. EGISD-YOLO: Edge Guidance Network for Infrared Ship Target Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 10097–10107. [[CrossRef](#)]
48. Wang, K.; Gao, X.; Zhao, Y.; Li, X.; Dou, D.; Xu, C.-Z. Pay attention to features, transfer learn faster CNNs. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.

**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.