Enhancing Infrared Small Target Detection: A Saliency-Guided Multi-Task Learning Approach

Zhaoying Liu, Yuxiang Zhang, Junran, He, Ting Zhang*, Sadaqat ur Rehman, Mohamad Saraee , Changming Sun

***Abstract*—Object detection in infrared images poses a considerable challenge due to its small-scale targets, low contrast and poor signal-to-clutter ratio, often resulting in a high false alarm rate. To improve the detection accuracy on infrared small targets, we introduce Light-SGMTLM, a lightweight and saliency-guided multi-task learning model. This model integrates saliency detection into the YOLOv5x framework through a parallel multi-task learning structure and employs a joint loss function during training. Such integration significantly alleviates the impact of complex backgrounds and improves the precision of small target localization. Moreover, we have developed a streamlined module, termed SIWD, to create a more agile backbone, which establishes an optimal balance between precision and efficiency, making the model more suitable for situations with limited computational resources. Comprehensive comparative experiments were conducted on six infrared small target datasets, namely, Small-ExtIRShip, Small-SSDD, IHAST, NUAA-SIRST, IRSTD-1k, and IRDST, and we assessed the model's performance against ten leading target detection models, such as YOLOv7, YOLOv8, DINO, and Relation-DETR. The findings reveal that our method's unique joint learning architecture, combining saliency and object detection tasks, significantly improves accuracy for infrared small target detection. Notably, it achieved impressive mean average precision (mAP) values of 92.60% and 75.71% on the NUAA-SIRST and IRSTD-1k datasets, respectively.**

***Index Terms*—Infrared small target detection, feature fusion, sailency detection, multi-task learning, lightweight.**

I. Introduction

INFRARED small target detection plays a vital role in a wide range of military and civilian applications including infrared guidance [1], early warning of natural disasters [2, 3], and maritime surveillance [4–7]. In comparison to visible imaging, detecting multiple small targets within a single infrared image frame constitutes a challenging endeavor due to their distinct characteristics. First, the infrared imaging's long-range characteristics result in an exceedingly small proportion of the target within the image, often consisting of just a few pixels or even a single pixel in extreme cases. Therefore, multilayer convolution and downsampling within the feature extraction process can easily cause the loss of small target features. Additionally, a significant imbalance arises between the positive and negative samples which are easily categorized mostly [8], so the model cannot be optimized in the desired direction. Second, owing to the distinctive imaging properties inherent to infrared imagery, pertinent information such as color, shape, and texture is commonly absent [9]. Finally, the background of an infrared small target is complex, contains substantial amounts of noise and exhibits a low signal-to-clutter ratio (SCR) [10]. Consequently, the target becomes prone to being overwhelmed by the background [11]. Thus, it is very difficult to detect small infrared targets using the network designed for normal objects [12–17].

Therefore, most of the existing research on infrared small target detection treats it as a semantic segmentation task [18–22]. An underlying reason for its effectiveness may stem from the semantic segmentation's ability to classify each pixel in the image, thereby mitigating the challenges associated with small target scales. However, segmentation tasks require meticulous processing of the intricacies of the pixel level, which involves considerable computational resources. Consequently, both the training and inference phases are commonly protracted. Furthermore, semantic segmentation serves as an intermediate representation used solely for tracking and localizing small infrared targets. The coherence of segmentation merely approximates the detection accuracy, while the precise detection performance remains unassessable. In summary, infrared small target detection based on general object detection methods typically yields inaccurate results. Conversely, segmentation-based approaches frequently incur extended computational time during the inference phase and struggle to address cases of overlapping targets. So it is necessary to design a robust and effective end-to-end infrared small targets detection method.

An intuitive idea is to integrate the object detection with the segmentation, leveraging the strengths of these two tasks to learn more comprehensive features of small targets. Recent studies have validated the advantages of multi-task learning [23–27], including improved data utilization efficiency, alleviation of overfitting, and enhancement of model performance. Prevailing method for constructing multi-task learning architectures involves incorporating a detection head into a segmentation model or embedding a segmentation head into a detection model. As for other parts they share parameters and features, namely hard parameter sharing (HPS). It is worth mentioning that compared to segmentation models, the backbone design of detection models is more cost-effective, as it can capture more non-local semantic information with the same number of downsampling and upsampling stages, which is beneficial for detecting small objects. Based on these

Zhaoying Liu, Yuxiang Zhang, Junran He, and Ting Zhang are with College of Computer Science, Beijing University of Technology, Beijing, 100124, China (e-mail: zhaoying.liu@bjut.edu.cn;)

S. Rehman is with the School of Sciences, Engineering and Environment, University of Salford, UK (email: s.rehman15@salford.ac.uk)

M. Saraee is with the School of Sciences, Engineering and Environment, University of Salford, UK (email: m.saraee@salford.ac.uk)

Changming Sun is with CSIRO Data61, PO Box 76, Epping, NSW 1710, Australia (e-mail: changming.sun@csiro.au).

*Corresponding author

email: zhangting@bjut.edu.cn

motivations and observations, we introduced a saliency guidance model based on YOLOv5x, and constructing a parallel multi-task learning structure using hard parameter sharing. Additionally, inspired by the concepts of dilated convolution and decomposed convolution [28, 29], we also designed a simple inception with dilation (SIWD) backbone to further reduce the number of model parameters and computational complexity.

In summary, our contributions are as follows:

- We propose a lightweight multi-task learning model, named Light-SGMTLM. By integrating saliency detection and employing multi-task learning techniques, this model significantly enhances the detection accuracy of existing object detection models for infrared images.
- We propose a saliency detection module that effectively reduces the impact of complex backgrounds. This module also facilitates the integration of deep and shallow features, further enhancing the accuracy of small target detection.
- We have designed a lightweight module named SIWD which is capable of significantly reducing the number of model parameters while maintaining the size of the receptive field. This module can be easily integrated into most detection models.
- The results of the comparative experiments with ten current state-of-the-art models, such as YOLOv5, Deformable DETR, DAB-DETR, DN-DETR, and DINO, on datasets including Small-ExtIRShip, Small-SSDD, IHAST, NUAA-SIRST, IRSTD-1k, and IRDST, demonstrate that our method outperforms other CNN-based models on the mAP metric and Transformer-based models on FLOPs metric.

II. Related Work

The categorization of infrared small target detection algorithms predominantly comprises two ways: sequence detection (Tracking Before Detection, TBD) [30], [31], and single-frame detection (Detection Before Tracking, DBT). As TBD algorithms necessitate the integration of multi-frame imagery and demonstrate suboptimal performance under real-time conditions, we exclusively focuses on single-frame methods. The current single-frame ISTD methods can be classified into two main categories: model-based traditional methods and data-driven deep learning methods. In the subsequent section, we will delve into a more comprehensive review of these methodologies.

A. Model-Based ISTD Method

The traditional methods for infrared small target detection involve directly constructing a model to quantify the distinction between an infrared small target and its surrounding environment, which typically comprise the following three categories: filter-based methods, human visual system (HVS)-based methods and low-rank sparse based methods.

The filter-based methods utilize specifically designed templates to filter the input images, and thereby facilitating the detection of small targets through effectively enhancing the intensity of targets. Shi et al. [32] introduced an upgraded version of the high-boost filter, which effectively preserved the high-frequency signals related to small targets while suppressing low-frequency background signals. Kong et al. [33] employed Haar wavelet decomposition to filter the original image, followed by the application of a wavelet energy fusion algorithm to isolate small targets. Zhang et al. [34] proposed a novel gradient correlation filtering (GCF) method, which was integrated with local consumption characteristics to achieve accurate distinction between small targets and clutter.

The human visual system suggests that the most salient regions in an image are determined not by brightness alone, but rather by the level of contrast they possess. Therefore they commonly rely on calculating the intensity difference between the target and its surrounding neighborhood for target identification. Chen et al. [35] proposed an LCM method that measured the contrast between the central and its neighboring region, which improved the intensity of small targets and suppressed clutters. Deng et al. [36] put forward a WLDM method, which assesses the local variance of each pixel across multiple scales, subsequently assigning modified local entropy as the weight for this variance.

The low-rank sparse methods distinguish the target and background by leveraging the sparsity of the target and the low-rank properties of the background. For example, Gao et al. [37] introduced the IPI model, utilizing a fixed-size window to traverse the entire image, extracting various patches, and subsequently vectorizing each patch for background estimation. To address the issue of inaccurate estimation of strong edges, Dai et al. [38] employed the nonnegative constraint and partial sum of singular values for background estimation. Gao et al. [39] proposed a graph learning-enhanced matrix decomposition method for separating cirrus clouds from backgrounds. By integrating this method with weighted local fractal features, they effectively suppressed strong edge noise and improved the detection capability of dim cirrus clouds.

These methods are computationally efficient, however, as they depends on the manual designed features and hyperparameters, high false alarm rate may occur for images with complex backgrounds.

B. Data-Driven ISTD Method

Over the past decade, deep learning methods have shown a remarkable capacity to dynamically extract image features and capture high-level semantic information. Consequently, these methods have outperformed traditional approaches in handling diverse complex environments. Furthermore, the availability of numerous infrared small target datasets has sparked growing interest among researchers in deep learning-based methods. For example, in order to enhance the detection performance of small infrared targets, Li and Shen [40] devised a technique that combines super-resolution enhancement of the input image with refinements to the structure of YOLOv5. In a comparable context, Zhou et al. [41] addressed the task of infrared small target detection through the utilization of a YOLO-based framework. Dai et al. [8] proposed a one-stage cascaded refinement network (OSCAR), which aims to

alleviate the inherent defects and inaccuracies of bounding box regression encountered in detecting small infrared targets. Meanwhile, Yao et al. [42] devised a lightweight network amalgamating conventional filtering methods with the standard FCOS framework to enhance the responsiveness towards infrared small target detection. To alleviate the issue of extreme foreground-background imbalance in infrared target detection tasks, Yang et al. proposed the adaptive threshold focal loss (ATFL), which utilizes an adaptive mechanism to adjust loss weights, compelling the detector to allocate more attention to foreground features.

In addition to the aforementioned object detection approaches, numerous studies have regarded ISTD as a semantic segmentation task which utilizes pixel-level threshold to yield a segmentation mask. Dai et al. [43] introduced an asymmetric contextual modulation (ACM) module, integrating top-down and bottom-up point-wise attention mechanisms to amplify the semantic information. Furthermore, Dai et al. [44] introduced an attentional local contrast network (ALCNet), to overcome the receptive field constraints and facilitating the interaction of long-range contextual information. To mitigate the loss of deep information induced by pooling layers in infrared small target detection, Li et al. [45] introduced the dense nested attention network (DNA-Net), which incorporates stacked U-shaped structures to extract significantfeatures. Additionally, Zhang et al. [46] proposed an attention-guided contextual module, which captures pixel correlations within and between blocks at different scales through local semantic association and global contextual attention. For real-time detection of infrared targets, Kou et al. [19] introduced the depth-wise separable atrous asymmetricatrous module (DAAA), which can effectively reduce computational complexity while learn multiscale features of small infrared targets.

Although these deep learning methods circumvent the need for extensive prior knowledge in manual design and achieve satisfactory performance, they are devised based on individual task. Therefore, there are certain limitations in the generalization ability of these methods.

C. Multi-task Learning

Multi-task learning aims to train multiple related tasks at the same time, achieve knowledge transfer by sharing underlying features, and promote the model to learn richer feature representations, thereby improving the performance and generalization capabilities of the model. Chen et al. [4] proposed a novel end-to-end framework for infrared small target detection and segmentation named MTUNet, and achieved a higher accuracy of infrared small target localization by incorporating a simple anchor-free detection head into the segmentation network. Xu et al. [24] proposed MTFormer, a Transformer-based multi-task learning architecture that integrates semantic segmentation, depth estimation, and saliency detection tasks. Additionally, they devised a cross-task attention mechanism to achieve adaptive sharing among different tasks. Experimental results demonstrate that this approach leads to performance improvements across all tasks. However, as tasks increases, the complexity of the decoder increases accordingly. To solve this problem, Xin et al. [47] integrated the decoder-free vision-language model CLIP with multi-task learning, proposing a tuning framework based on Multi-modal Alignment Prompts (MmAP). By combining the zero-shot generalization capabilities of CLIP with MmAP, this framework enhances information interaction and sharing between tasks, thereby effectively improving the performance and efficiency of multi-task learning.

Although these methods improve the performance of each subtask, most of them are targeted at visible scenes, how to use multi-task learning to improve the location ability of infrared small targets is still a topic worth exploring. Therefore, we try to construct multi-task learning based on saliency detection task and object detection task in infrared scene, so as to improve the detection accuracy of infrared small targets.

III. Proposed Method

A. Overall Architecture

The workflow of our proposed method is illustrated in Fig. 1. Given an infrared image as input, features across five different scales are acquired through a lightweight backbone comprised of a stack of five SIWD modules. Subsequently, we utilize the FPN [48] to fuse features from the latter three stages, thus integrating multi-scale information. Following this, the upsampling stage within the proposed saliency detection module is utilized to further integrate shallow local features extracted from the first and second stages with the output of FPN. Simultaneously, while feeding the fused features into the saliency detection head, the downsampling stage is employed to pass shallow information of small targets back to the PAN [49]. Finally, leveraging the saliency detection head and the detection head of YOLOv5x separately for saliency detection and bounding box prediction, we construct a joint loss function and utilize multi-task learning for end-to-end training.

B. Saliency Detection Module

Due to the longer distances involved in infrared imaging, most targets appear as small objects. However, current object detectors generally only rely on deeply aggregated semantic information for prediction, with feature maps having much lower resolutions than the original input images. Although operations such as FPN and PAN are employed for feature fusion, they are limited to relatively deeper features and do not consider shallow features containing rich detailed information, which are very important for small object detection. To mitigate this issue, we expanded the depth of FPN and PAN by introducing a saliency detection module as shown in Fig. 1.

The proposed saliency detection module comprises two components: feature extraction and saliency detection head. The feature extraction module comprises upsampling fusion and downsampling fusion. Specifically, we incorporate cross-stage residual blocks (CSPF) to the FPN and PAN, respectively. The CSPFs in upsampling fusion process aim to alleviate the aliasing effect resulting from the fusion of upsampling features and lateral connection features, while facilitating

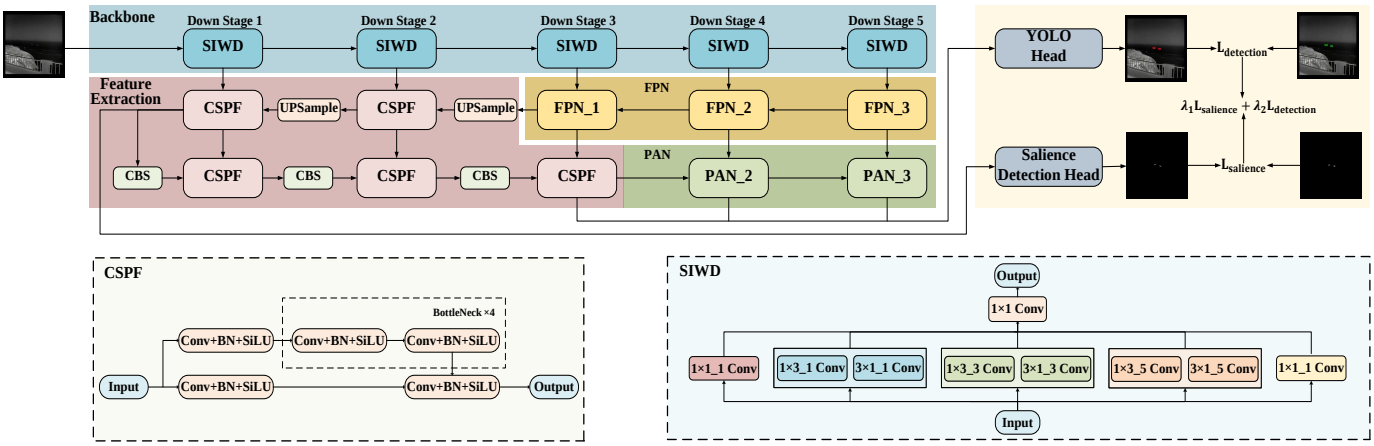

Fig. 1. The overall structure of the proposed Light-SGMTLM.

the fusion of deep semantic information and shallow local features. The CSPFs in downsampling fusion process aims to pass the small target information back to PAN. Furthermore, to alleviate the mitigating issues related to information loss and gradient vanishing, we discarded the skip connections present in the original CSP in our proposed CSPF module, as shown in Fig. 2. The fusion feature map of up-sampling is obtained via:

$$FF_i = \begin{cases} O_i\left(X_i \odot Up\left(Y\right)\right) & i = 2 \\ O_i\left(X_i \odot Up\left(FF_{i+1}\right)\right) & i = 1 \end{cases} \tag{1}$$

where FF_i represents the output features of each layer of the up-sampling fusion, and O_i represents the i-th CSPF operation, and i decreases from the deep layer to the shallow layer. X_i is the feature map of the *i*-th output of the first two Down Stages (Down Stage 1 and Down Stage 2). Y represents the output of the last layer of FPN. Up represents the up-sampling operation based on bilinear interpolation. $\odot$ represents channel-based feature fusion. Meanwhile, the down-sampling process is shown in (2) :

$$FP_i = \begin{cases} O_i\left(FF_i \odot CBS\left(FF_i\right)\right) & i = 1, 2 \\ O_i\left(FF_i \odot CBS\left(Y\right)\right) & i = 3 \end{cases} \tag{2}$$

where FP_i represents the output features of each layer of the down-sampling fusion, and CBS indicates the combination operations of convolution, BN, and SiLU.

The saliency detection head consists of two convolution operations and a sigmoid function. The process of obtaining the final saliency map is as follows:

$$S = \sigma\left(\mathrm{Conv}\,7\left(\mathrm{Conv}\,6\left(R\left(F_{\mathrm{up}}\right)\right)\right)\right) \tag{3}$$

where Conv6 and Conv7 are convolution operations for dimensionality reduction, and the size of the convolution kernel is 1×1. R represents the CSPS module, which shares the same structure as the CSPF. σ represents a sigmoid operation, which is used to compute the probability of each pixel and predict whether it belongs to the saliency target, so as to obtain the saliency feature map S.

C. Lightweight Backbone Feature Extraction Network

While the incorporation of a saliency detection module for multi-task learning amplifies the feature information of small targets and suppresses background noise, it also introduces increased model complexity, which imposes limitations on its applicability in practical scenarios with constraints on memory and computational resources. To alleviate the issue, we designed a simple inception module named SIWD. Inspired by the Inception structure [50], we proposed a multi-branch architecture. Within it, the conventional convolutions are decomposed and substituted using the principles of asymmetric and dilated convolutions. The SIWD structure employs asymmetric convolutions with varying dilation rates to ensure a substantial reduction in parameter count while maintaining a consistent receptive field. The following sections provide detailed descriptions of the backbone network and the SIWD module.

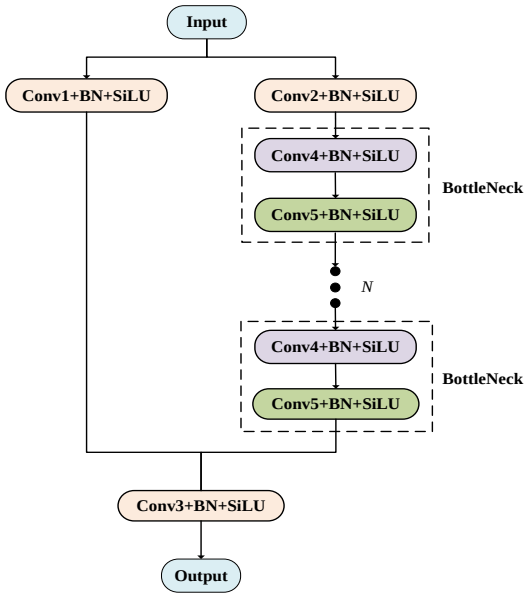

Fig. 2. The structure of the CSPF module, where *N* is set to 4.

From Fig. 1, it can be observed that the main component of the entire model comprises five down-sampling stages, each of which corresponds to an SIWD module. These modules are equipped with convolutional kernels of varying sizes and

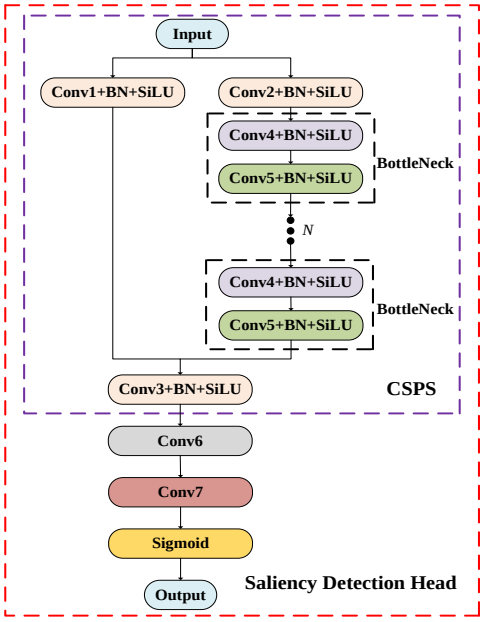

Fig. 3. The structure of the saliency detection head. The inside purple box is the residual structure of CSPS.

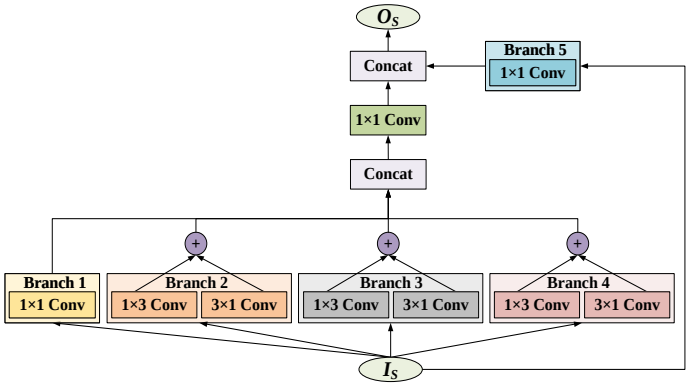

Fig. 4. The structure diagram of the SIWD module. The layered architecture of the SIWD module highlights its key components such as dilated convolution layers and factorized convolution units. It illustrates how the module efficiently processes infrared images, reducing computational complexity while maintaining accuracy for small target detection in complex backgrounds.

TABLE I
PARAMETER CONFIGURATION OF SIWD IN EACH DOWN-SAMPLING STAGE.

Module	Stage	Input Channel	Output Channel
SIWD1	Down Stage 1	3	80
SIWD2	Down Stage 2	80	160
SIWD3	Down Stage 3	160	320
SIWD4	Down Stage 4	320	640
SIWD5	Down Stage 5	640	1,280

dilation rates, and the changes in channel numbers before and after each down-sampling stage are illustrated in Table I. To reduce the model's parameter count, the SIWD module replaces the 3×3 convolutional kernels with 1×3 and 3×1 convolutional kernels and employs dilated convolutions to expand the receptive field. As depicted in Fig. 4, the SIWD module primarily consists of five branches. The first branch comprises a 1×1 convolutional layer with a dilation rate of 1, aimed at dimension reduction. The second branch employs two convolutional layers with kernel sizes of 3×1 and 1×3, both with a dilation rate of 1. The input feature maps are individually processed by these two convolutional layers and then pointwise added to form the output of this branch. The structures of the third and fourth branches are similar to the second branch, but with dilation rates of 3 and 5, respectively, to extract features with different receptive field sizes. The fifth branch draws inspiration from the residual network's Shortcut mechanism, with input features undergoing dimension reduction through 1×1 convolutions to obtain output features. After each of the input feature maps has traversed these five branches, the outputs from the first four branches are concatenated, followed by fusion through a pointwise convolution layer to yield O_L. Finally, this result is concatenated with the output from the fifth branch to obtain the final output O_S of the SIWD module. The overall processing flow of the SIWD module can be defined as follows:

$$O_S = \left(C^1_{1\times1}\left(B_1 \odot B_2 \odot B_3 \odot B_4\right)\right) \odot B_5 \tag{4}$$

$$B_i = \begin{cases} A_{\text{add}}\left(C^{2i-3}_{1\times3}\left(I_S\right), C^{2i-3}_{3\times1}\left(I_S\right)\right) & i = 2,3,4 \\ A_{\text{add}}\left(C^1_{1\times1}\left(I_S\right), C^1_{1\times1}\left(I_S\right)\right) & i = 1,5 \end{cases} \tag{5}$$

where C^y_x denotes the combination of a convolution layer with kernel size of x and dilation rate of y, a batch normalization layer, and a SiLU activation operation. A_{add} denotes the point-by-point summation of the feature maps, $\odot$ denotes the splicing of the feature maps, and B_i denotes the output feature maps of the i-th branch.

Assuming that the dimensions of the input and output features of the SIWD module are denoted as $C_{\text{in}} \times H_{\text{in}} \times W_{\text{in}}$ and $C_{\text{out}} \times H_{\text{out}} \times W_{\text{out}}$, respectively, the size of the output feature maps for the first four branches is $\frac{1}{4}C_{\text{out}} \times \frac{1}{2}H_{\text{in}} \times \frac{1}{2}W_{\text{in}}$, while the size of the output feature maps for the fifth branch and after the pointwise convolution is $\frac{1}{2}C_{\text{out}} \times \frac{1}{2}H_{\text{in}} \times \frac{1}{2}W_{\text{in}}$. As a convolutional kernel with dimensions $k \times 1$ and a dilation rate of d is essentially equivalent to a convolution of size $((d-1) \times 1 + k) \times 1$, and a convolutional kernel with dimensions $1 \times k$ is essentially equivalent to a convolution of size $1 \times (1 \times (d-1) + k)$, so the parameter count for each branch of SIWD can be defined as in (6), where P_i represents the parameter count for the i-th branch. The total parameter count for the SIWD module is the summation of the parameter counts for its five branches and the pointwise convolution, denoted as $(\frac{21}{4}C_{\text{in}} + \frac{1}{2}C_{\text{out}}) \times C_{\text{out}}$.

$$P_i = \begin{cases} 1 \times 1 \times C_{\text{in}} \times \frac{1}{4}C_{\text{out}} & i = 1 \\ (1 \times 3 \times C_{\text{in}} + 3 \times 1 \times C_{\text{in}}) \times \frac{1}{4}C_{\text{out}} & i = 2,3,4 \\ 1 \times 1 \times C_{\text{in}} \times \frac{1}{2}C_{\text{out}} & i = 5 \end{cases} \tag{6}$$

When not employing dilated convolutions in the decomposition of convolutional layers, only the parameters of the second,

third, and fourth branches differ from the rest, as illustrated in (7). The parameter count remains constant for the other branches. Therefore, the total parameter count without using dilated asymmetric convolution decomposition is $(21C_{\text{in}} + \frac{1}{2}C_{\text{out}}) \times C_{\text{out}}$. The difference between the two is $15.75C_{\text{in}}C_{\text{out}}$. This can be a substantial reduction in the aggregated parameter count, indicating a noteworthy reduction.

$$P_i^* = \left((2i-1)^2 \times C_{\text{in}}\right) \times \frac{1}{4}C_{\text{out}} \quad (i = 2, 3, 4) \tag{7}$$

D. Joint Multi-task Loss Function

The loss function of Light-SGMTLM proposed in this paper consists of two parts: saliency loss and object detection loss, as shown in (8), where λ_1 and λ_2 are the weights of sailency loss and target detection loss, respectively. We conducted ablation experiments on the proportion settings of these two parameters and selected the appropriate weights.

$$L_{\text{total}} = \lambda_1 L_{\text{salience}} + \lambda_2 L_{\text{det}} \tag{8}$$

where the saliency loss L_{saliency} consists of the cross-entropy loss L_{ce} and the structural similarity loss L_{ssim} as shown in (9). L_{ce} denotes the similarity measure between the final predicted significance score map and the real label, while L_{ssim} denotes the overall similarity measure between the two in terms of structural features including brightness and contrast.

$$L_{\text{saliency}} = L_{\text{ce}} + (1 - L_{\text{ssim}}) \tag{9}$$

The infrared images typically encompass crucial structural information, including the contour, texture, and shape of the target. By incorporating a structural similarity loss as an auxiliary loss for saliency detection tasks, the model can be forced to learn the inherent texture and shape features of the target within the image. This facilitates better alignment with the original image's structure and generates saliency region images of higher quality. Consequently, it reduces sensitivity to image noise and environmental variations while reducing false alarm rates. Combining this approach with cross-entropy loss helps in improving convergence speed. They are defined below:

$$L_{\text{ce}} = -\sum_{i=1}^{N} \left(Y_i \log\left(S_i\right) + \left(1 - Y_i\right) \log\left(1 - S_i\right)\right) \tag{10}$$

$$L_{\text{ssim}} = \frac{\left(2\mu_Y \mu_S + k_1\right)\left(2\sigma_{YS} + k_2\right)}{\left(\mu_Y^2 + \mu_S^2 + k_1\right)\left(\sigma_Y^2 + \sigma_S^2 + k_2\right)} \tag{11}$$

Among them, Y and S represent the prediction results of the real label and saliency detection respectively, N represents the number of pixels, μ_Y and μ_S are the mean values of Y and S respectively, σ_Y and σ_S are the standard deviations of Y and S respectively, σ_{YS} is the covariance of Y and S, k_1 and k_2 are two constants used to prevent the dividend from being 0, and they were set to 0.001 and 0.0009 respectively in the experiment.

$$L_{\text{det}} = L_{\text{reg}} + L_{\text{conf}} + L_{\text{cls}} \tag{12}$$

The target detection loss consists of bounding box regression loss L_{reg}, confidence loss L_{conf}, and classification loss L_{cls}, as shown in (12). For the bounding box regression loss, as given in (13), λ_{coord} represents the weight of the positive sample, K represents the number of grids divided into each row and column of the input image, and M represents the preset anchor in each grid. For the number of boxes, M is set to 3 in this paper, and the value of I_{ij}^{obj} is 0 or 1, which indicates whether the j-th bounding box in the i-th grid is used to predict the target. w_i and h_i represent the width and height of the prediction box respectively, and D represents the area of the smallest circumscribed rectangle between the prediction box and the real box.

$$L_{\text{reg}} = \lambda_{\text{coord}} \sum_{i=1}^{K \times K} \sum_{j=1}^{M} I_{ij}^{\text{obj}} L_{\text{GIoU}} \left(2 - w_i \times h_i\right) \tag{13}$$

$$L_{\text{GIoU}} = 1 - GIoU = 1 - IoU + \frac{|D - A \cup B|}{|D|} \tag{14}$$

For the confidence loss, as shown in (15), λ_{noobj} represents the weight of the negative sample. I_{ij}^{noobj} indicates whether the j-th prediction box of the i-th grid is responsible for predicting the target. If it is not responsible for predicting the target, it takes 1, otherwise it takes 0. $\bar{C}_i$ indicates whether there is a target in the i-th grid. If it is, it takes 1, otherwise it takes 0.

$$\begin{aligned} L_{\text{conf}} = &- \sum_{i=1}^{K \times K} \sum_{j=1}^{M} I_{ij}^{\text{obj}} \left[\bar{C}_i \log\left(C_i\right) + \left(1 - \bar{C}_i\right) \log\left(1 - C_i\right)\right] \\ &- \lambda_{\text{noobj}} \sum_{i=1}^{K \times K} \sum_{j=1}^{M} I_{ij}^{\text{noobj}} \left[\bar{C}_i \log\left(C_i\right) + \left(1 - \bar{C}_i\right) \log\left(1 - C_i\right)\right] \end{aligned} \tag{15}$$

For the classification loss, as shown in (16), $\bar{P}_i(n)$ represents the true value of category n in the i-th grid. If it belongs to category n , its size is 1, otherwise it is 0, and $P_i(n)$ represents the predicted probability of category in the i-th grid.

$$\begin{aligned} L_{\text{cls}} = &- \sum_{i=1}^{K \times K} \sum_{j=1}^{M} I_{ij}^{\text{obj}} \\ &\sum_{c \in classes} \left[\bar{P}_i(n) \log(P_i(n)) + (1 - \bar{P}_i(n)) \log(1 - P_i(n))\right] \end{aligned} \tag{16}$$

IV. Experiment

A. Experimental Details

1) Implementation Settings: Our experiments were conducted on a computer equipped with an NVIDIA Tesla K40c graphics card and the software environment utilized the PyTorch framework, with Python version 3.6. During training, we configured the initial learning rate to be 0.001, employed a batch size of 8, and utilized the Adam optimizer with a weight decay of 0.0005 for a total of 50 epochs.

2) Datasets: We curated and annotated the saliency detection datasets in GL-Light-NLDF [51] and the publicly available dataset SSDD [52]. Subsequently, two multi-task infrared ship target detection datasets were constructed, namely Small-ExtIRShip and Small-SSDD. Additionally, four public datasets were also utilized: IRDST [53], IHAST [54], IRSTD-1k [55], and NUAA-SIRST [43].

The Small-ExtIRShip dataset is derived from the saliency detection dataset GL-Light-NLDF. Firstly, we employed a relative proportion method to screen small targets with a pixel proportion less than 0.12%, resulting in 1,270 small target images. Secondly, since the ExtIRShip dataset lacks target detection labels, we annotated the target box to obtain the multi-task Small-ExtIRShip dataset. Finally, we divided the training and test sets at a ratio of 9:1, resulting in 1,143 training images and 127 test images.

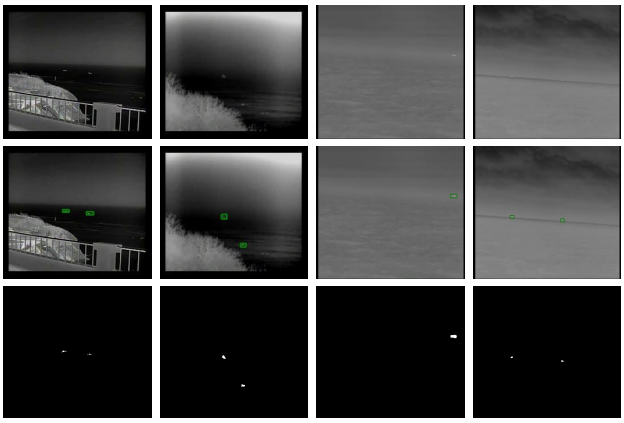

Fig. 5. Example images from the Small-ExtIRShip dataset. The first row represents the original images while the second and third rows depict the object detection and saliency detection labels respectively.

The Small-SSDD dataset is derived from the publicly available SSDD dataset, which already includes object detection and saliency labels. Hence, it requires filtering the dataset based on the relative proportion of pixels less than 0.12% and subsequently dividing it into a training set and a test set in a 9:1 ratio, resulting in 252 images for training and 28 images for testing.

The IHAST dataset consists of 1,200 infrared images featuring complex, dynamic backgrounds and high-speed small targets, categorized into aircraft, drones, and birds. We annotated the dataset with pixel-level masks to support multi-task model training. The dataset is split into training and test sets in a 7:3 ratio.

The IRDST dataset contains 142,727 images, including 40,650 real images from 85 scenes and 102,077 simulated images from 317 scenes. Each image is annotated with pixel-level masks, bounding boxes, and central pixels. In this study, only the 40,650 real images were used, split into training and test sets at a 7:3 ratio, resulting in 28,460 training images and 12,196 test images.

The NUAA-SIRST dataset includes 427 images selected from infrared video sequences. However, the original target detection labels were inaccurate, often grouping multiple targets within an image under a single label. We corrected this by re-annotating individual bounding boxes using mask-based modifications. The final dataset, containing around 480 instances, is split into training and test sets at a 7:3 ratio.

The IRSTD-1k dataset comprises a total of 1,000 infrared images. It is already provided with the regression labels and

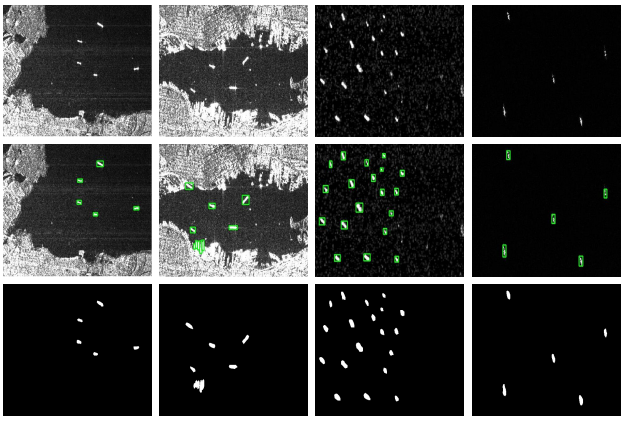

Fig. 6. Example images from the Small-SSDD dataset. The first row represents the original images while the second and third rows depict the object detection and saliency detection labels respectively.

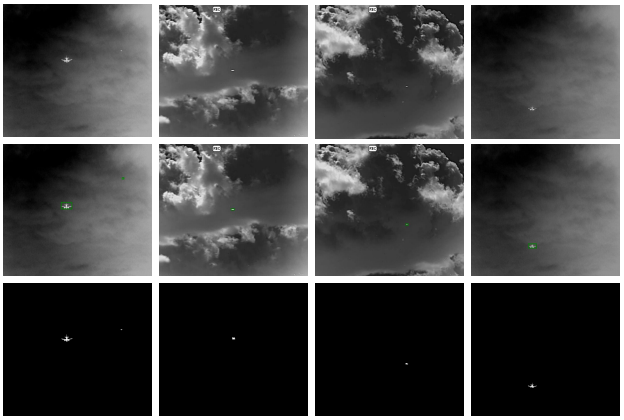

Fig. 7. Example images from the IHAST dataset. The first row represents the original images while the second and third rows depict the object detection and saliency detection labels respectively.

Fig. 8. Example images from the IRDST dataset. The first row represents the original images while the second and third rows depict the object detection and saliency detection labels respectively.

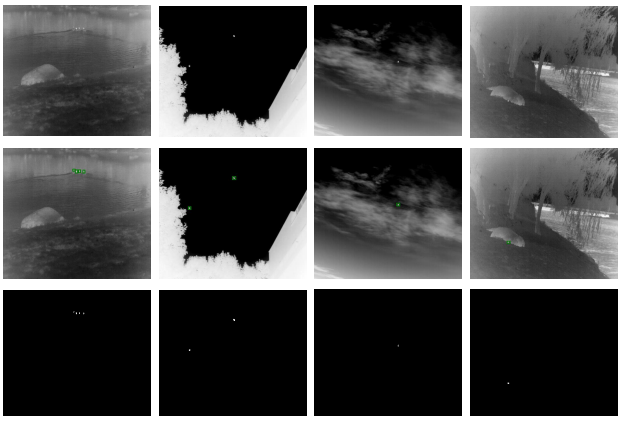

Fig. 9. Example images from the NUAA-SIRST dataset. The first row represents the original images while the second and third rows depict the object detection and saliency detection labels respectively.

pixel-level masks, thereby requiring only the division of the training set and test set in a 7:3 ratio. Data augmentation involves random flipping and adjustments to image brightness and contrast while no data augmentation was used during testing. Some examples of these six datasets are demonstrated in Fig. 5 to Fig. 10. Throughout all the comparison experiments and ablation experiments conducted in this study, fixed input sizes were employed during training: the Small-ExtIRShip and Small-SSDD datasets had an input size of 416×416, the NUAA-SIRST dataset utilized an input size of 320×320, the IHAST dataset adopted an input size of 608×608, the IRSTD-1k dataset featured an input size of 512×512, and the IRSTD dataset employed an input size of 992×992.

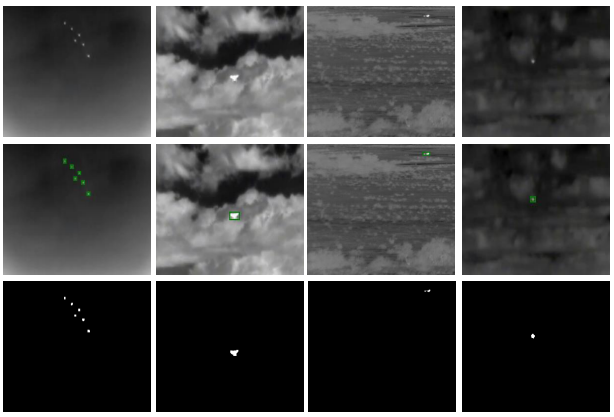

Fig. 10. Example images from the IRSTD-1k dataset. The first row represents the original images while the second and third rows depict the object detection and saliency detection labels respectively.

3) Evaluation Metrics: We adopt commonly used evaluation metrics including multi-class average precision (mAP) , number of model parameters (Params) , floating point operations (Flops) , intersection over union (IoU) and normalized intersection over union (nIoU). The calculations of mAP refers to the average AP value across multiple classes, which is calculated as the area under the curve, with recall on the x-axis and precision on the y-axis, ranging from 0 to 1. While typically computed via integration, in practice, an approximation is often used by summing the precision at each threshold, multiplied by the corresponding change in recall.

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$mAP = \frac{\sum_{i=1}^{M} AP(i)}{M} \tag{19}$$

IoU aims to measure the accuracy of detecting the corresponding object in a given dataset, i.e.,

$$\text{IoU} = A_i / A_u \tag{20}$$

where A_i and A_u are intersection region and union region between the prediction and ground truth, respectively.

nIoU is the normalization of IoU. It can make a better balance between structural similarity and pixel accuracy of infrared small target and is defined as follows:

$$\text{nIoU} = \frac{1}{N} \sum_{i=1}^{N} (\text{TP}[i]/(T[i] + P[i] - \text{TP}[i])) \tag{21}$$

where N represents the total number of the target pixels and $TP[\cdot]$ denotes the number of true positive pixels.

B. Comparative Experiments of Object Detection

To validate the effectiveness of the proposed Light-SGMTLM, we conducted comparative experiments against fourteen state-of-the-art detection models, i.e., YOLOv3 [56] , YOLOv4 [49] , YOLOv5x [57] , YOLOv6l [13] , YOLOv7x [58] , YOLOv8x [59] , Deformable DETR [14] , DAB-DETR [15] , DN-DETR [16] , DINO [17] , Rank-DETR [60] , MS-DETR[61] , Salience-DETR [62] and Relation-DETR [63]. The experimental results are presented in Table II. From the results, it is evident that our method achieves a significantly higher detection accuracy compared to convolution-based approaches. Additionally, compared to state-of-the-art transformer-based methods, our approach demonstrates superior performance in terms of both parameter efficiency and computational cost.

Fig. 11 and Fig. 12 present the visual results of different comparison methods against complex sky and ocean backgrounds, respectively. From the visualization results, it is evident that our method is more effective in suppressing false alarms at cloud boundaries, identifying small targets hidden within cloud layers, and exhibiting greater robustness against high-intensity sea wave noise near coastlines compared to other methods.

TABLE II

THE COMPARATIVE EXPERIMENTAL RESULTS OF MULTIPLE OBJECT DETECTION MODELS, WHERE DM-DETR DENOTES DEFROMABLE DETR. THE UPPER SECTION OF THE TABLE CORRESPONDS TO MODELS BASED ON CNNS, WHILE THE LOWER SECTION PERTAINS TO MODELS BASED ON TRANSFORMERS. THE METRICS MARKED IN RED AND BLUE CORRESPOND TO THE BEST AND SECOND PERFORMANCE RESPECTIVELY.

Model	Params (M)	Small-ExtIRShip		Small-SSDD		NUAA-SIRST		IHAST		IRSTD-1k		IRDST	
		FLOPs (G)	mAP	FLOPs (G)	mAP	FLOPs (G)	mAP	FLOPs (G)	mAP	FLOPs (G)	mAP	FLOPs (G)	mAP
YOLOv3	61.53	32.80	86.95	32.80	86.35	19.41	74.16	70.07	81.33	49.69	63.04	186.53	74.74
YOLOv4	63.94	29.98	88.15	29.98	88.21	17.74	75.65	64.04	82.24	45.42	64.62	170.48	75.82
YOLOv5x	83.21	45.91	92.27	45.91	92.60	54.45	78.34	196.58	85.53	139.41	67.73	523.30	75.93
YOLOv6l	59.60	32.22	93.48	32.22	94.32	19.06	78.56	68.81	86.18	48.80	68.92	183.17	76.03
YOLOv7x	71.75	42.48	94.53	42.48	95.68	25.14	81.25	90.74	86.97	64.35	70.42	241.56	76.84
YOLOv8x	91.90	48.47	95.10	48.47	96.10	28.68	82.30	103.52	88.90	73.42	72.53	275.57	77.43
Deformable-DETR	41.02	78.14	92.51	78.14	93.22	46.23	78.90	166.90	85.70	118.36	63.44	444.30	76.30
DAB-DETR	44.38	84.54	93.82	84.54	93.80	50.02	81.33	180.57	86.50	128.05	64.64	480.69	76.10
DN-DETR	47.31	90.12	94.41	90.12	94.90	53.32	84.56	192.49	88.50	136.51	66.81	512.43	77.50
DINO	47.78	91.01	94.72	91.01	95.20	53.85	87.70	194.4	91.80	137.86	70.80	517.52	78.30
Rank-DETR	49.83	95.12	94.33	95.12	95.36	56.29	88.46	203.20	92.10	144.10	72.44	540.92	78.33
MS-DETR	51.21	105.69	94.71	105.69	95.92	62.54	88.90	225.78	92.22	160.11	73.84	601.02	78.38
Salience-DETR	159.60	104.99	94.89	104.99	96.01	60.29	89.15	216.26	92.46	154.33	74.68	579.34	78.42
Relation-DETR	153.64	123.31	**95.72**	123.31	**96.41**	72.96	**90.28**	250.01	**92.88**	177.29	**74.92**	665.54	**78.55**
Light-SGMTLM	45.53	36.51	**96.66**	36.51	**97.43**	19.68	**92.60**	71.05	**93.40**	50.39	**75.71**	189.16	**78.92**

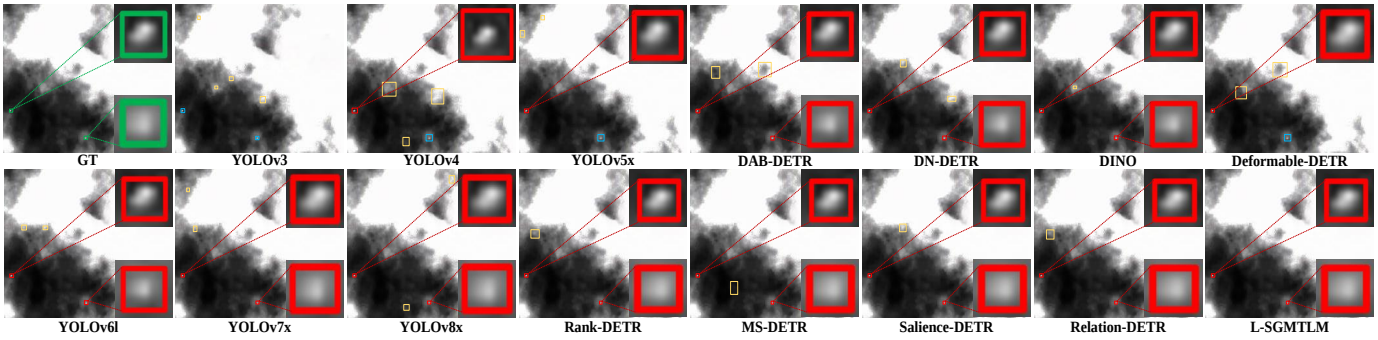

Fig. 11. Detection results under complex cloud background. The red, yellow and blue boxes represent correctly predicted targets, false alarms and missed targets, while the green boxes denote the ground truth.

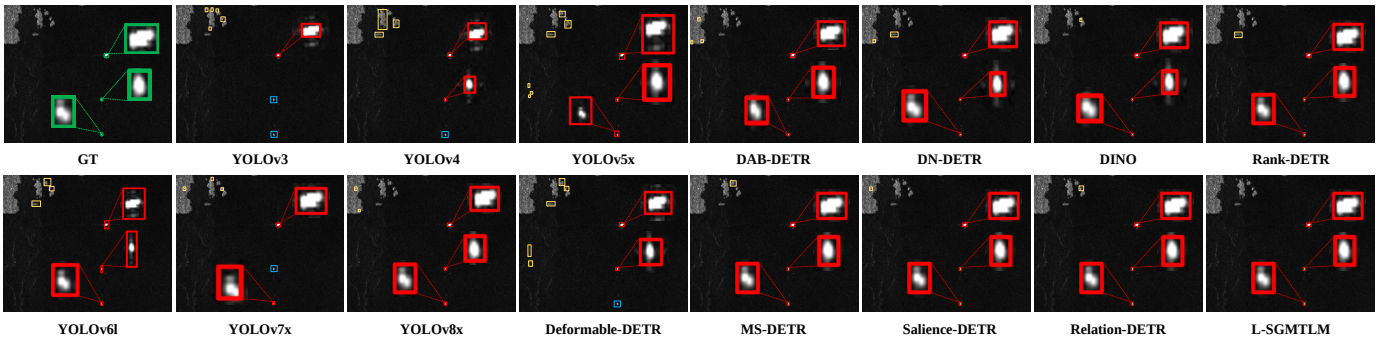

Fig. 12. Detection results in a near-shore marine background. The red, yellow and blue boxes represent correctly predicted targets, false alarms and missed targets, while the green boxes denote the ground truth.

TABLE III
THE COMPARATIVE EXPERIMENTAL RESULTS OF MULTIPLE SEGMENTATION MODELS, THE METRICS MARKED IN RED AND BLUE CORRESPOND TO THE BEST AND SECOND PERFORMANCE RESPECTIVELY.

Model	NUAA-SIRST		IRSTD-1k		IHAST		ExtIRShip		Small-SSDD	
	IoU(%)	nIoU(%)	IoU(%)	nIoU(%)	IoU(%)	nIoU(%)	IoU(%)	nIoU(%)	IoU(%)	nIoU(%)
ISNet	68.78	70.73	61.21	59.18	75.04	65.35	62.32	61.76	73.94	71.09
ISTDU	63.34	62.58	51.58	50.26	70.01	69.84	51.88	51.02	72.61	70.87
FC3Net	64.64	64.49	53.08	50.43	72.87	66.31	53.88	53.79	74.85	70.64
IAANet	66.43	67.83	54.66	53.89	72.64	72.02	54.63	53.78	73.24	69.58
HCFNet	68.33	69.28	56.48	55.80	75.86	**73.51**	60.92	59.10	74.83	72.96
DNANet	67.90	68.09	64.53	**62.54**	77.82	69.17	**64.15**	**64.34**	73.31	69.26
SCTransNet	69.64	68.99	64.34	59.29	78.07	71.27	56.46	56.70	78.52	75.73
MiMNet	**69.92**	**70.80**	**65.02**	61.30	**78.31**	72.98	57.30	56.64	**78.58**	**75.85**
Light-SGMTLM	**70.61**	**71.03**	**65.22**	**63.37**	**79.50**	**73.67**	**65.78**	**63.45**	**78.67**	**76.18**

C. Comparative Experiments of Saliency Detection

To further demonstrate the effectiveness of our proposed method, we also compare it with another eight advanced methods designed for ISTD, including ISNet [64], ISTDU [65], FC3Net [66], IAANet [67], HCFNet [68], DNANet [69], SCTransNet [70] and MiMNet [71]. The quantitative results are presented in Table III. It can be seen from Table III that our proposed method demonstrated superior results with the optimal or suboptimal IoU and nIoU metrics across multiple datasets. Notably, our method achieves an IoU of 65.78% on the ExtIRShip dataset, surpassing the runner-up by 1.63%. Some visualization examples of results have shown in Fig. 13 and Fig. 14, including multiple targets under homogeneous cloud and water surface with high-brightness reflection noise.

As shown in Fig. 13, due to the lack of color and texture, extremely small targets in the clouds generally exhibit homogeneity with the background, leading to the phenomenon of missed detection. Whereas, our method can effectively identify these targets hidden in clouds. Fig. 14 demonstrates that our proposed method can also suppress false alarms caused by coastal and wave reflections effectively. Additionally, compared to other methods, the segmentation results generated by our approach closely approximate the real shapes, further validating the effectiveness in capturing infrared small target features.

D. Loss Function Weights Analysis

The influence of the loss weight λ_1 for the saliency detection task and the loss weight λ_2 for the target detection task on the accuracy of infrared small target detection in the multi-task loss function is presented in Table IV. The results indicate that when the target detection loss weight exceeds that of saliency detection, the model tends to prioritize target detection, negatively affecting both saliency detection and small target localization. However, with $\lambda_1 = 1$ and λ_2=1, a balanced approach is achieved, leading to higher accuracy in small target detection. Therefore, setting both weights to 1 is recommended.

TABLE IV
WEIGHTED ANALYSIS EXPERIMENTAL RESULTS OF MULTI-TASK LOSS FUNCTION ON INFRARED SMALL TARGET DATASETS.

λ_1	λ_2	mAP				
		Small-ExtIRShip	Small-SSDD	NUAA-SIRST	IHAST	IRSTD-1k
1	2	95.42	96.96	91.08	90.94	75.03
2	1	95.35	96.41	92.56	91.41	75.08
1	3	96.21	96.47	91.94	92.21	73.16
3	1	95.39	96.22	91.34	91.39	73.05
2	3	91.68	97.30	90.60	92.84	72.79
3	2	91.74	97.07	91.39	92.04	74.88
1	4	90.53	97.01	90.97	93.10	75.51
4	1	90.87	97.23	90.89	92.22	69.35
1	5	90.67	97.60	90.75	92.77	71.70
5	1	91.40	**98.32**	90.44	92.43	72.80
1	1	**96.66**	97.43	**92.60**	**93.40**	**75.71**

E. Effectiveness Analysis of Each Structure of the Saliency Detection Module

We conducted ablation experiments on the Small-ExtIRShip, Small-SSDD, and IHAST datasets to evaluate the feature extraction and saliency detection head in our proposed saliency detection branch, keeping other network components unchanged. As shown in Table V, the addition of feature extraction improved detection accuracy on small targets by 1.64%, 3.12%, and 4.12% on the three datasets, respectively. When both feature extraction and saliency detection heads

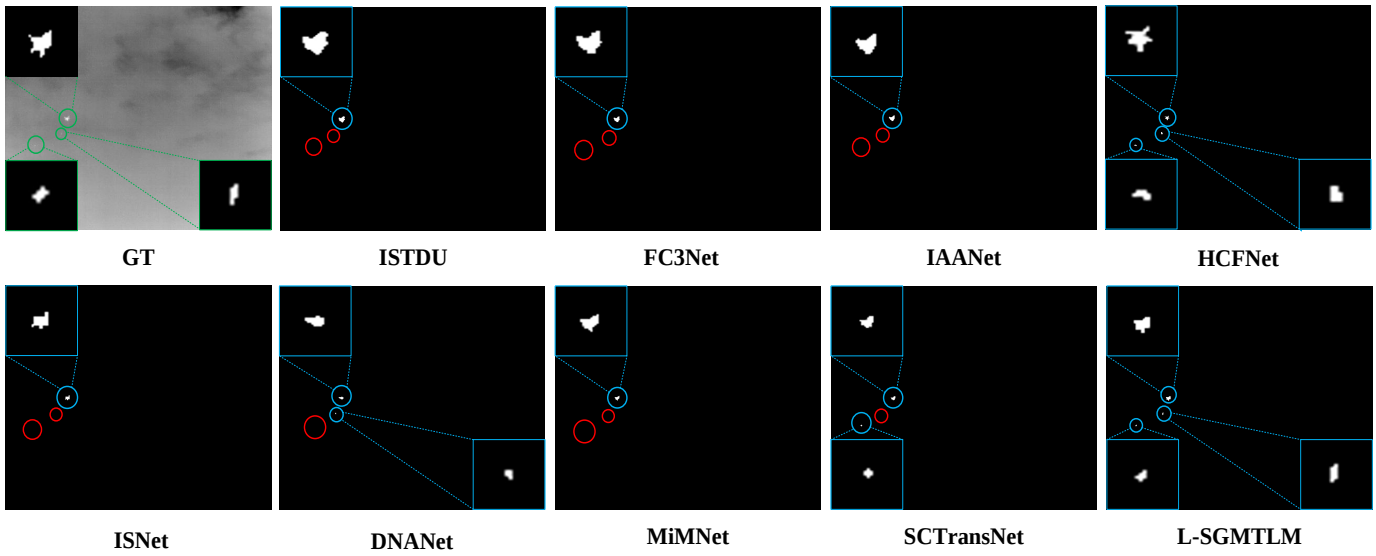

Fig. 13. Segmentation results obtained by different ISTD methods under a large number of homogeneous cloud backgrounds. The blue and red circles represent correctly predicted and missed detection respectively, while the green boxes denote ground truth.

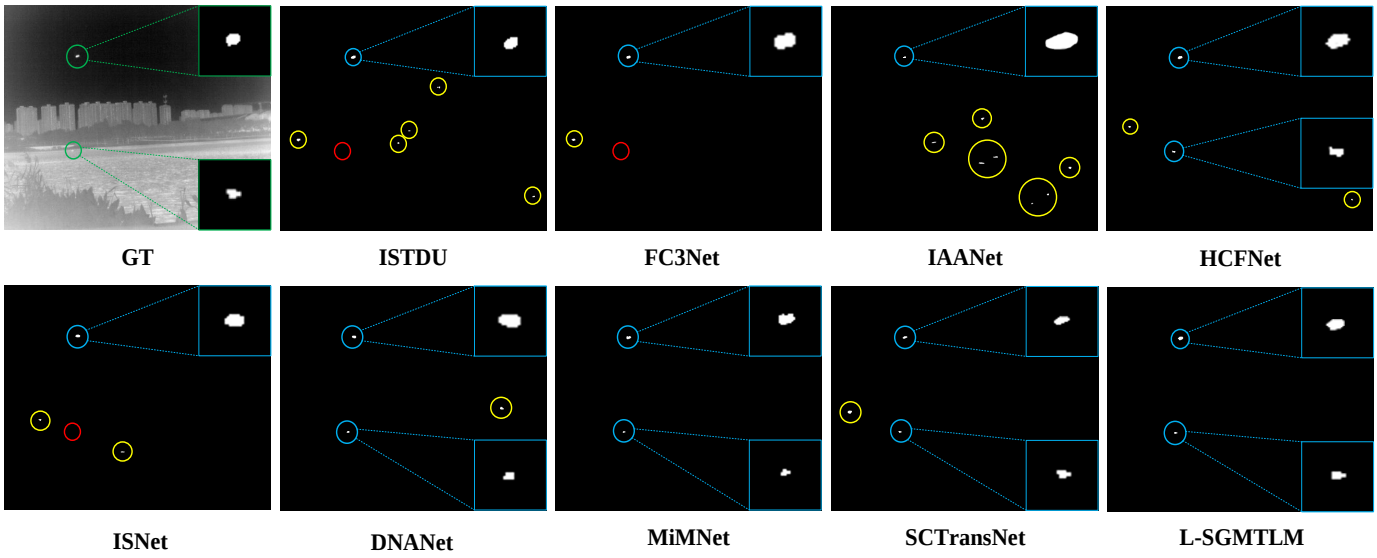

Fig. 14. Segmentation results obtained by different ISTD methods under the background of high-brightness reflection noise on the water surface. The blue, yellow and red circles represent correctly predicted, false alarm and missed detection respectively, while the green boxes denote ground truth.

were included, accuracy further increased by 3%, 5.93%, and 7.92%. These results demonstrate that feature extraction enhances the fusion of shallow details with deep semantic features, enriching the input to the detection head and boosting accuracy. Additionally, the saliency detection task further improves infrared small target detection. To visualize the impact, we used heatmaps on the Small-SSDD dataset, comparing the original YOLOv5x model, YOLOv5x* (with feature extraction), and our L-SGMTLM model (with both modules). As shown in Fig. 15, feature extraction improves detection by integrating local and semantic information, though it can also introduce non-target features. However, adding saliency detection effectively suppresses background noise, guiding the model to focus on small targets and improving overall accuracy.

TABLE V
THE RESULTS OF ABLATION EXPERIMENTS ON FEATURE EXTRACTION AND SALIENCY DETECTION HEAD ON INFRARED SMALL TARGET DATASETS.

Feature Extraction	Saliency Detection Head	mAP		
		Small-ExtIRShip	Small-SSDD	IHAST
×	×	93.66	91.50	85.48
✓	×	95.30	94.62	89.60
✓	✓	**96.66**	**97.43**	**93.40**

F. Analysis of the Applicability of Each Structure in the Saliency Detection Module

The proposed Light-SGMTLM is developed based on YOLOv5x. To further validate the effectiveness of the feature

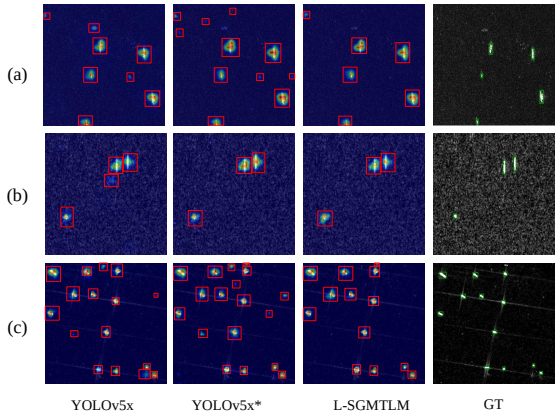

Fig. 15. The heatmap of the Small-SSDD dataset, where (a), (b) and (c) respectively represent the detection results of three pictures in the Small-SSDD dataset under the original YOLOv5x, YOLOv5x after adding the feature fusion part, and Light-SGMTLM model, as well as their corresponding ground truth.

fusion component and saliency detection header in the saliency detection module, we utilize different versions of YOLOv5 as backbones, incorporating feature extraction and saliency detection headers separately. Subsequently, experiments are conducted on the Small-ExtIRShip and Small-SSDD datasets to assess its applicability. The experimental results are presented in Table VI.

The result reveals that the feature extraction component and the saliency detection head within YOLOv5s, YOLOv5m, and YOLOv5l exhibit equal applicability for infrared small target detection tasks, demonstrating the generalizability of the saliency detection module across these YOLOv5 models.

TABLE VI
THE ABLATION EXPERIMENTAL RESULTS OF THE FRAMEWORK BASED ON SALIENCY GUIDANCED MULTI-TASK LEARNING ON THE SMALL-EXTIRSHIP AND SMALL-SSDD DATASETS.

Backbone	Feature Extraction	Saliency Detection Head	mAP	
			Small-ExtIRShip	Small-SSDD
YOLOv5s	×	×	89.19	87.52
	✓	×	90.72	88.61
	✓	✓	**93.74**	**89.67**
YOLOv5m	×	×	90.19	88.41
	✓	×	92.11	89.38
	✓	✓	**94.37**	**91.62**
YOLOv5l	×	×	91.96	90.80
	✓	×	93.35	92.61
	✓	✓	**95.41**	**93.29**

G. Validation Analysis of the SIWD Module

To assess the impact of the proposed lightweight module SIWD on model performance, we first conducted ablation experiments on the Small-ExtIRShip and Small-SSDD datasets with various scales of YOLOv5 and our proposed Light-SGMTLM. Results are presented in Table VII. From the results, it is evident that our proposed SIWD module, for infrared small target detection tasks, significantly reduces model parameters, floating-point operations, and cumulative operations while attempting to preserve model detection accuracy.

Moreover, to further verify the transferability of SIWD, we integrated the SIWD structure with different resource intensive models including DANet [72] , SegFormer [73] , SETR-PUP [74] , kMaX-DeepLab [75] and OneFormer [76]. Experiments were conducted on the NUAA-SIRST and IRSTD-1k datasets and the results are shown in Table VIII and Table IX. From the results, it is evident that the proposed SIWD module can be seamlessly integrated into various large models, reducing their parameter count and computational complexity while simultaneously improving segmentation accuracy. This enhancement improves the applicability of the model in scenarios with limited computational resources.

TABLE VII
ANALYSIS OF THE IMPACT OF THE SIWD MODULES ON THE SMALL-EXTIRSHIP AND SMALL-SSDD DATASETS.

Models	SIWD	Param (M)	FLOPs (G)	MAdd (G)	mAP	
					Small-ExtIRShip	Small-SSDD
YOLOv5s	×	6.91	4.83	9.65	**89.19**	**87.52**
	✓	**4.19**	**3.15**	**6.27**	88.73	86.94
YOLOv5m	×	20.62	14.96	29.89	**90.19**	88.41
	✓	**11.73**	**9.08**	**18.11**	89.68	**90.02**
YOLOv5l	×	45.68	34.0	67.94	**91.96**	90.80
	✓	**25.03**	**20.98**	**41.88**	91.57	**92.31**
YOLOv5x	×	83.21	45.91	91.81	92.27	**92.60**
	✓	**43.30**	**20.11**	**40.18**	**94.19**	91.50
Light-SGMTLM	×	85.44	62.32	124.55	**96.82**	**97.66**
	✓	**45.53**	**36.51**	**72.92**	96.66	97.43

TABLE VIII
TRANSFERABILITY EVALUATION OF THE SIWD MODULE ON THE NUAA-SIRST DATASET.

Models	Backbone	Param (M)	FLOPs (G)	IoU(%)	nIoU(%)
DANet	ResNet101	66.60	**126.86**	64.98	62.88
	SWID	**43.48**	133.24	**66.79**	**64.25**
SegFormer	Mit-B5	84.59	38.92	66.95	65.01
	SWID	**4.77**	**17.32**	**67.30**	**66.38**
SETR-PUP	ViT-Large	318.3	228.72	65.64	63.29
	SWID	**33.62**	**113.15**	**65.92**	**63.53**
kMaX-DeepLab	ConvNeXt-L	232.0	84.25	68.58	67.02
	SWID	**47.92**	**17.25**	**68.76**	**67.33**
OneFormer	DiNAT-L	223.0	90.75	69.18	68.20
	SWID	**35.78**	**23.32**	**70.29**	**69.34**

H. Comparative Analysis of Different Lightweight Modules

To investigate the impact of using different lightweight modules to replace the backbone feature network on the performance of the lightweight multi-task model, we employ InceptionV3 [50], ShuffleNet2 [77], SqueezeNet1 [78], and

TABLE IX
TRANSFERABILITY EVALUATION OF THE SIWD MODULE ON THE IRSTD-1K DATASET.

Models	Backbone	Param (M)	FLOPs (G)	IoU(%)	nIoU(%)
DANet	ResNet101	66.60	**324.76**	58.63	54.41
	SWID	**43.48**	338.46	**60.76**	**55.53**
SegFormer	Mit-B5	84.59	99.64	61.87	55.32
	SWID	**4.77**	**44.35**	**62.16**	**56.73**
SETR-PUP	ViT-Large	318.3	577.78	60.73	54.61
	SWID	**33.62**	**288.90**	**60.78**	**55.16**
kMaX-DeepLab	ConvNeXt-L	232.0	213.12	62.54	59.45
	SWID	**47.92**	**41.44**	**62.71**	**59.66**
OneFormer	DiNAT-L	223.0	229.76	62.78	60.14
	SWID	**35.78**	**56.92**	**62.89**	**59.72**

SIWD as replacements for the backbone feature extraction network in this section. The parameters of each lightweight module are adjusted to match the input dimensions at each downsampling stage of Light-SGMTLM while maintaining other parts unchanged. Comparative experiments are conducted on the Small-SSDD, IRSTD-1k, NUAA-SIRST, and IHAST datasets using these lightweight modules. The first dataset primarily consists of marine ship scenes, characterized by significant wave noise interference, and the scenes in the rest datasets predominantly consist of sky and urban, accompanied by significant interference from cloud noise and complex background. Moreover, the input sizes of these four datasets exhibit inconsistency. The experimental results are presented in Table XI and Table XIII.

The results indicate that when InceptionV3 is employed as the backbone network, it exhibits the highest number of parameters, floating-point computation, and MAdd. This can be attributed to its larger network depth and width, along with the adoption of multiple stacked lightweight modules for enhanced feature extraction, resulting in a higher parameter count. On the other hand, SIWD is derived from Inception but only incorporates an SIWD module structure in the five down-sampling stages within Light-SGMTLM. Consequently, when SIWD serves as the feature extraction network instead of InceptionV3, there is a reduction in parameter count, floating-point computation, and MAdd for the model. Remarkably, this replacement of backbone network with the SIWD module leads to significant reductions across different datasets and input sizes without compromising detection accuracy. In fact, improvements may even occur. Notably although the parameter count of SIWD module exceeds SqueezeNet1 by 6.57M, our proposed model demonstrates lower floating-point computation and MAdd on all datasets while achieving higher detection accuracy. These findings highlight that among various lightweight modules considered herein, SIWD outperforms others within lightweight multi-task models.

TABLE X
THE COMPARATIVE EXPERIMENTS OF DIFFERENT LIGHTWEIGHT MODULES OF LIGHT-SGMTLM ON THE SMALL-SSDD AND IRSTD-1K DATASETS.

Backbone	Param (M)	Small-SSDD			IRSTD-1k		
		FLOPs (G)	MAdd (G)	mAP	FLOPs (G)	MAdd (G)	mAP
InceptionV3	47.96	64.99	129.81	95.62	98.44	196.64	69.52
ShuffleNetV2	38.72	39.93	79.74	94.37	60.49	120.79	69.77
SqueezeNetV1	**38.96**	37.68	75.20	94.66	57.07	113.91	68.89
SIWD	45.53	**36.51**	**72.92**	**97.43**	**50.39**	**100.62**	**75.71**

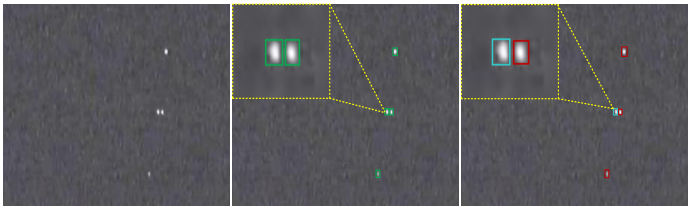

Fig. 16. Example of missed detection when detecting closely targets, where the red, blue and green boxes represent missed detection, correct detection and ground truth respectively.

I. Limitations and Future Work

Although our method demonstrates good performance on the six datasets, it still exhibits instances of missed detection when confronted with overlapping or dense small targets, as illustrated in Fig. 16. It is caused by the sensitivity of Intersection over Union (IoU) metrics concerning small targets, resulting in some detection boxes being filtered out during post-processing by Non-Maximum Suppression (NMS).

In addition, our model requires high-dimensional feature maps to capture rich semantic information, which results in a high number of parameters. Therefore, we attempt to reduce the dimensions of the feature maps. Specifically, we reduce the number of channels of the output feature maps of each stage in the backbone to one-tenth of their original number, while keeping other parts unchanged. Then, we conduct experiments on six datasets, including Small-ExtIRShip, Small-SSDD, NUAA-SIRST, IRSTD-1k, IHAST, and IRDST. The results are shown in the Table XI and Table XIII. It can be observed that although dimension reduction effectively decreases the number of parameters, it significantly slows down the convergence speed. After reducing the number of channels, it often takes 600 to 1200 epochs to converge, whereas previously, only 50 to 70 epochs were needed to achieve convergence. The extended training time increases the risk of over-fitting and reduces the generalization ability of the model. Therefore, in this work, we chose to maintain the number of channels in backbone.

In future work, we will focus on enhancing NMS to deal with small dense target with overlapping and to further lightweight the model while ensuring training speed and accuracy.

V. CONCLUSION

In this paper, our contributions encompass two main aspects: firstly, we propose a multi-task learning model for infrared

TABLE XI
COMPARATIVE RESULTS ON THE SMALL-SSDD, IRSTD-1K AND IHAST DATASETS BEFORE AND AFTER REDUCING THE NUMBER OF CHANNELS. WHERE †REPRESENTS THE MODEL AFTER REDUCING THE NUMBER OF CHANNELS.

Model	Param (M)	Small-SSDD			IRSTD-1k			IHAST		
		FLOPs (G)	mAP	Epoch	FLOPs (G)	mAP	Epoch	FLOPs (G)	mAP	Epoch
Light-SGMTLM†	0.49	0.43	97.27	700	0.65	74.94	700	0.92	92.66	700
Light-SGMTLM	45.53	36.51	97.43	50	50.39	75.71	70	71.05	93.40	50

TABLE XII
COMPARATIVE RESULTS ON THE NUAA-SIRST, SMALL-EXTIRSHIP AND IRDST DATASETS BEFORE AND AFTER REDUCING THE NUMBER OF CHANNELS. WHERE †REPRESENTS THE MODEL AFTER REDUCING THE NUMBER OF CHANNELS.

Model	Param (M)	NUAA-SIRST			Small-ExtIRShip			IRDST		
		FLOPs (G)	mAP	Epoch	FLOPs (G)	mAP	Epoch	FLOPs (G)	mAP	Epoch
Light-SGMTLM†	0.49	0.26	91.38	1200	0.43	92.33	800	2.46	75.08	900
Light-SGMTLM	45.53	19.68	92.60	60	36.51	96.66	50	189.16	78.92	50

TABLE XIII
THE COMPARATIVE EXPERIMENTS OF DIFFERENT LIGHTWEIGHT MODULES OF LIGHT-SGMTLM ON NUAA-SIRST AND IHAST DATASETS.

Backbone	Param (M)	NUAA-SIRST			IHAST		
		FLOPs (G)	MAdd (G)	mAP	FLOPs (G)	MAdd (G)	mAP
InceptionV3	47.96	38.45	76.81	90.12	138.82	277.29	91.49
ShuffleNetV2	38.72	23.63	47.18	89.52	85.30	170.34	90.29
SqueezeNetV1	**38.96**	22.29	44.50	91.27	80.48	160.64	90.35
SIWD	45.53	**19.68**	**39.31**	**92.60**	**71.05**	**141.90**	**93.40**

small target detection by aggregating a saliency detection module with YOLOv5x. Specifically, it consists of two components, namely feature extraction and saliency detection head. The feature extraction section integrates shallow and deep information from neck to capture more comprehensive feature sets and then divided into two branches: one portion is directed towards the saliency detection head to generate saliency map and participate in the calculation of the joint loss, while the other is passed back to the neck to guide the PAN to focus on the small target area. Secondly, we devised a streamlined implant module based on factorized convolution and dilated convolution, termed SIWD, to establish an agile backbone network. In our experiments, we conducted extensive evaluations on six datasets and compared our approach with some state-of-the-art methods. The experimental results substantiate that our proposed method not only adeptly enhances the performance of infrared small target detection, thereby mitigating false alarms, but also notably diminishes parameter count and computational complexity.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (62176009, U23A20357, 61806013, 61906005), General Project of Science and Technology Plan of Beijing Municipal Education Commission (KM202110005028), International Research Cooperation Seed Fund of Beijing University of Technology (2021A01), and Project of Interdisciplinary Research Institute of Beijing University of Technology (2021020101).

REFERENCES

[1] X. Tong, S. Su, P. Wu, R. Guo, J. Wei, Z. Zuo, and B. Sun, "MSAFFNet: A Multiscale Label-Supervised Attention Feature Fusion Network for Infrared Small Target Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[2] S. Bai, L. Yang, Y. Liu, and H. Yu, "Dmf-net: A dual-encoding multi-scale fusion network for pavement crack detection," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–16, 2023.

[3] S. ur Rehman, Z. Yang, M. Shahid, N. Wei, Y. Huang, M. Waqas, S. Tu, and O. U. Rehman, "Water preservation in soan river basin using deep learning techniques," *arXiv: 1906.10852*, 2019.

[4] Y. Chen, L. Li, X. Liu, and X. Su, "A multi-task framework for infrared small target detection and segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–9, 2022.

[5] X. Li, T. Zhang, Z. Liu, B. Liu, S. ur Rehman, B. Rehman, and C. Sun, "Saliency guided siamese attention network for infrared ship target tracking," *IEEE Transactions on Intelligent Vehicles*, 2024.

[6] T. Zhang, G. Jiang, Z. Liu, S. ur Rehman, and Y. Li, "Advanced integrated segmentation approach for semi-supervised infrared ship target identification," *Alexandria Engineering Journal*, vol. 87, pp. 17–30, 2024.

[7] T. Zhang, H. Shen, S. ur Rehman, Z. Liu, Y. Li, and O. ur Rehman, "Two-stage domain adaptation for infrared ship target segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[8] Y. Dai, X. Li, F. Zhou, Y. Qian, Y. Chen, and J. Yang, "One-stage cascade refinement networks for infrared

small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[9] D. Zhou and X. Wang, "Research on high robust infrared small target detection method in complex background," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[10] R. Kou, C. Wang, Y. Yu, Z. Peng, F. Huang, and Q. Fu, "Infrared small target tracking algorithm via segmentation network and multi-strategy fusion," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[11] C. Li, Y. Zhang, Z. Shi, Y. Zhang, and Y. Zhang, "Moderately dense adaptive feature fusion network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 7464–7475.

[13] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[14] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," 2021.

[15] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=oMI9PjOb9Jl

[16] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR Training by Introducing Query DeNoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 13 619–13 627.

[17] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection," 2022.

[18] X. Ying, L. Liu, Y. Wang, R. Li, N. Chen, Z. Lin, W. Sheng, and S. Zhou, "Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 15 528–15 538.

[19] R. Kou, C. Wang, Y. Yu, Z. Peng, M. Yang, F. Huang, and Q. Fu, "Lw-irstnet: Lightweight infrared small target segmentation network and application deployment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[20] S. ur Rehman, Y. Huang, S. Tu, and B. Ahmad, "Learning a semantic space for modeling images, tags and feelings in cross-media search," in *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2019 Workshops, BDM, DLKT, LDRC, PAISI, WeL, Macau, China, April 14–17, 2019, Revised Selected Papers 23*. Springer, 2019, pp. 65–76.

[21] B. Nian, B. Jiang, H. Shi, and Y. Zhang, "Local contrast attention guide network for detecting infrared small targets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[22] T. Guo, B. Zhou, F. Luo, L. Zhang, and X. Gao, "Dmfnet: Dual-encoder multistage feature fusion network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[23] W. Sheng and X. Li, "Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network," *Pattern Recognition*, vol. 114, p. 107868, 2021.

[24] X. Xu, H. Zhao, V. Vineet, S.-N. Lim, and A. Torralba, "MTFormer: Multi-task learning via transformer and cross-task reasoning," in *European Conference on Computer Vision*. Springer, 2022, pp. 304–321.

[25] H. Tan, T. Ye, S. ur Rehman, O. ur Rehman, S. Tu, and J. Ahmad, "A novel routing optimization strategy based on reinforcement learning in perception layer networks," *Computer Networks*, vol. 237, p. 110105, 2023.

[26] V. Sampath, I. Maurtua, J. J. A. Martín, A. Rivera, J. Molina, and A. Gutierrez, "Attention guided multi-task learning for surface defect identification," *IEEE Transactions on Industrial Informatics*, 2023.

[27] D. N. Goncalves, J. M. Junior, P. Zamboni, H. Pistori, J. Li, K. Nogueira, and W. N. Goncalves, "MTLSegFormer: Multi-task Learning with Transformers for Semantic Segmentation in Precision Agriculture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6289–6297.

[28] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[29] Q. Qiu, X. Cheng, G. Sapiro *et al.*, "DCFNet: Deep neural network with decomposed convolutional filters," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4198–4207.

[30] F. Wu, H. Yu, A. Liu, J. Luo, and Z. Peng, "Infrared small target detection using spatiotemporal 4-d tensor train and ring unfolding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2023.

[31] J. Li, P. Zhang, L. Zhang, and Z. Zhang, "Sparse regularization-based spatial–temporal twist tensor model for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[32] Y. Shi, Y. Wei, H. Yao, D. Pan, and G. Xiao, "High-boost-based multiscale local contrast measure for infrared small target detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 1, pp. 33–37, 2017.

[33] X. Kong, L. Liu, Y. Qian, and M. Cui, "Automatic detection of sea-sky horizon line and small targets in maritime infrared imagery," *Infrared Physics & Technology*, vol. 76, pp. 185–199, 2016.

[34] X. Zhang, J. Ru, and C. Wu, "Infrared small target detection based on gradient correlation filtering and contrast measurement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[35] C. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE transactions on geoscience and remote sensing*, vol. 52, no. 1, pp. 574–581, 2013.

[36] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Small infrared target detection based on weighted local difference measure," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 4204–4214, 2016.

[37] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE transactions on image processing*, vol. 22, no. 12, pp. 4996–5009, 2013.

[38] Y. Dai, Y. Wu, Y. Song, and J. Guo, "Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values," *Infrared Physics & Technology*, vol. 81, pp. 182–194, 2017.

[39] Z. Gao, J. Yin, J. Luo, W. Li, and Z. Peng, "Multi-directional graph learning-based infrared cirrus detection with local texture features," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[40] R. Li and Y. Shen, "Yolosr-ist: A deep learning method for small target detection in infrared remote sensing images based on super-resolution and yolo," *Signal Processing*, vol. 208, p. 108962, 2023.

[41] X. Zhou, L. Jiang, C. Hu, S. Lei, T. Zhang, and X. Mou, "Yolo-sase: An improved yolo algorithm for the small targets detection in complex backgrounds," *Sensors*, vol. 22, no. 12, p. 4600, 2022.

[42] S. Yao, Q. Zhu, T. Zhang, W. Cui, and P. Yan, "Infrared image small-target detection based on improved fcos and spatio-temporal features," *Electronics*, vol. 11, no. 6, p. 933, 2022.

[43] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 950–959.

[44] Dai, Yimian and Wu, Yiquan and Zhou, Fei and Barnard, Kobus, "Attentional local contrast networks for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9813–9824, 2021.

[45] B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, and Y. Guo, "Dense nested attention network for infrared small target detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 1745–1758, 2023.

[46] T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-guided pyramid context networks for detecting infrared small target under complex background," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 4, pp. 4250–4261, 2023.

[47] Y. Xin, J. Du, Q. Wang, K. Yan, and S. Ding, "Mmap: Multi-modal alignment prompt for cross-domain multi-task learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 16 076–16 084.

[48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[49] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[51] Z. Liu, X. Zhang, T. Jiang, T. Zhang, B. Liu, M. Waqas, and Y. Li, "Infrared salient object detection based on global guided lightweight non-local deep features," *Infrared Physics & Technology*, vol. 115, p. 103672, 2021.

[52] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*. IEEE, 2017, pp. 1–6.

[53] H. Sun, J. Bai, F. Yang, and X. Bai, "Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset irdst," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[54] S. Xu, T. Zhang, Z. Liu, and Y. Li, "Effective infrared small target detection based on improved refinedet," in *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2021, pp. 103–107.

[55] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 877–886.

[56] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[57] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2778–2788.

[58] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464–7475.

[59] F. M. Talaat and H. ZainEldin, "An improved fire detection approach based on yolo-v8 for smart cities," *Neural Computing and Applications*, vol. 35, no. 28, pp. 20 939–20 954, 2023.

[60] Y. Pu, W. Liang, Y. Hao, Y. YUAN, Y. Yang, C. Zhang, H. Hu, and G. Huang, "Rank-detr for high quality object detection," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 16 100–16 113. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/

34074479ee2186a9f236b8fd03635372-Paper-Conference.pdf

[61] C. Zhao, Y. Sun, W. Wang, Q. Chen, E. Ding, Y. Yang, and J. Wang, "Ms-detr: Efficient detr training with mixed supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 17 027–17 036.

[62] X. Hou, M. Liu, S. Zhang, P. Wei, and B. Chen, "Salience detr: Enhancing detection transformer with hierarchical salience filtering refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 17 574–17 583.

[63] X. Hou, M. Liu, S. Zhang, P. Wei, B. Chen, and X. Lan, "Relation detr: Exploring explicit position relation prior for object detection," 2024. [Online]. Available: https://arxiv.org/abs/2407.11699

[64] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "Isnet: Shape matters for infrared small target detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 877–886.

[65] Q. Hou, L. Zhang, F. Tan, Y. Xi, H. Zheng, and N. Li, "Istdu-net: Infrared small-target detection u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[66] M. Zhang, K. Yue, J. Zhang, Y. Li, and X. Gao, "Exploring feature compensation and cross-level correlation for infrared small target detection," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1857–1865.

[67] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[68] S. Xu, S. Zheng, W. Xu, R. Xu, C. Wang, J. Zhang, X. Teng, A. Li, and L. Guo, "Hcf-net: Hierarchical context fusion network for infrared small object detection," *arXiv preprint arXiv:2403.10778*, 2024.

[69] B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, and Y. Guo, "Dense nested attention network for infrared small target detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 1745–1758, 2022.

[70] S. Yuan, H. Qin, X. Yan, N. AKhtar, and A. Mian, "Sctransnet: Spatial-channel cross transformer network for infrared small target detection," *arXiv preprint arXiv:2401.15583*, 2024.

[71] T. Chen, Z. Tan, T. Gong, Q. Chu, Y. Wu, B. Liu, J. Ye, and N. Yu, "Mim-istd: Mamba-in-mamba for efficient infrared small target detection," *arXiv preprint arXiv:2403.02148*, 2024.

[72] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[73] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 12 077–12 090. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf

[74] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 6881–6890.

[75] Q. Yu, H. Wang, S. Qiao, M. Collins, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "kmax-deeplab: k-means mask transformer," 2023. [Online]. Available: https://arxiv.org/abs/2207.04044

[76] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 2989–2998.

[77] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[78] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and ¡0.5MB model size," 2016.

Zhaoying Liu was born in 1986. She received her PhD from Beihang University in 2015. From 2015 to 2017 she was a postdoctoral researcher at Beijing University of Technology (BJUT). Her main research interests include image processing, pattern recognition, and deep learning. She is currently an Associate Professor in the College of Computer Science, BJUT. She is a Member of China Computer Federation (CCF) and Chinese Association for Artificial Intelligence (CAAI).

Yuxiang Zhang was born in 1999. He is currently pursuing the M.S. degree with the College of Computer Science, Beijing University of Technology, Beijing, China. His main research interests include deep learning and computer vision.

Junran He was born in 1995. He once pursued the M.S. degree with the College of Computer Science, Beijing University of Technology, Beijing, China, and obtained the degree in 2023. His main research interests include deep learning and computer vision.

Ting Zhang was born in 1986. She is currently an Associate Professor in the College of Computer Science, Beijing University of Technology (BJUT). She received her PhD from BJUT in 2018. From 2018 to 2020 she was a postdoctoral researcher at BJUT. She is a member of China Computer Federation (CCF). Her current research interests include pattern recognition, deep learning, and natural language processing.

Sadaqat ur Rehman (M'18) received his PhD degree (Sept. 2015 – Jun. 2019) with the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, China. He is currently working as an Assistant Professor in AI with the School of Science, Engineering and Environment, University of Salford, UK. He has produced a world leading research activity in the fields of data science, machine learning, intelligent systems (with emphasis on artificial neural networks), semantic multimedia analysis, optimization and affective computing.

Mohamad Saraee is a Chair in Data Science, IEEE member, and Head of the Computer Science Department at the University of Salford. He is the founder and director of the DSAI Hub, with over 25 years of expertise in Machine Learning, Data/Text Mining, and Big Data Analytics. Prof. Saraee earned his Ph.D. from the University of Manchester and has published over 170 articles in high-tier journals and international conferences. He has led research projects exceeding £1.8M, collaborating with Innovate UK, the NHS, and industry partners. He is also the Editor-in-Chief of the International Journal of Web Research and has supervised 24 Ph.D. and MPhil students.

Changming Sun received the Ph.D. degree in computer vision from the Imperial College London, London, U.K., in 1992. He then joined CSIRO, Sydney, Australia, where he is currently the Principal Research Scientist carrying out research and working on applied projects. He is also a Conjoint Professor with the School of Computer Science and Engineering, University of New South Wales. His current research interests include computer vision, image analysis, and pattern recognition. He has served on the program/organizing committees of various international conferences. He is an Associate Editor of the EURASIP Journal on Image and Video Processing.