Position-Aware Relation Learning for RGB-Thermal Salient Object Detection

Heng Zhou, Chunna Tian, *Member, IEEE*, Zhenxi Zhang, Chengyang Li, Yuxuan Ding, Yongqiang Xie, and Zhongbo Li

***Abstract*— Salient object detection (SOD) is an important task in computer vision that aims to identify visually conspicuous regions in images. RGB-Thermal SOD combines two spectra to achieve better segmentation results. However, most existing methods for RGB-T SOD use boundary maps to learn sharp boundaries, which lead to sub-optimal performance as they ignore the interactions between isolated boundary pixels and other confident pixels. To address this issue, we propose a novel position-aware relation learning network (PRLNet) for RGB-T SOD. PRLNet explores the distance and direction relationships between pixels by designing an auxiliary task and optimizing the feature structure to strengthen intra-class compactness and inter-class separation. Our method consists of two main components: A signed distance map auxiliary module (SDMAM), and a feature refinement approach with direction field (FRDF). SDMAM improves the encoder feature representation by considering the distance relationship between foreground-background pixels and boundaries, which increases the inter-class separation between foreground and background features. FRDF rectifies the features of boundary neighborhoods by exploiting the features inside salient objects. It utilizes the direction relationship of object pixels to enhance the intra-class compactness of salient features. In addition, we constitute a transformer-based decoder to decode multispectral feature representation. Experimental results on three public RGB-T SOD datasets demonstrate that our proposed method not only outperforms the state-of-the-art methods, but also can be integrated with different backbone networks in a plug-and-play manner. Ablation study and visualizations further prove the validity and interpretability of our method.**

***Index Terms*— Salient object detection, RGB-thermal images, swin transformer, position-aware relation learning.**

Manuscript received 11 October 2022; revised 9 March 2023 and 16 April 2023; accepted 21 April 2023. Date of publication 1 May 2023; date of current version 4 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62173265, in part by the Fundamental Research Funds for the Central Universities, and in part by the Innovation Fund of Xidian University. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiwen Lu. *(Corresponding authors: Chunna Tian; Yongqiang Xie.)*

Heng Zhou is with the School of Electronic Engineering, Xidian University, Xi'an 710071, China, and also with the Institute of Systems Engineering, AMS, Beijing 100141, China (e-mail: hengzhou@stu.xidian.edu.cn).

Chunna Tian, Zhenxi Zhang, and Yuxuan Ding are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: chnatian@xidian.edu.cn; zxzhang_5@stu.xidian.edu.cn; yxding@stu.xidian.edu.cn).

Chengyang Li is with the School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China, and also with the Institute of Systems Engineering, AMS, Beijing 100141, China (e-mail: chengyang_li@stu.pku.edu.cn).

Yongqiang Xie and Zhongbo Li are with the Institute of Systems Engineering, AMS, Beijing 100141, China (e-mail: yqxie.ams@gmail.com; zbli83@foxmail.com).

Digital Object Identifier 10.1109/TIP.2023.3270801

I. Introduction

SALIENT object detection (SOD) is to segment the main conspicuous objects in the image at the pixel level by simulating the human visual system. In applications of image quality assessment [1], [2], image editing [3], [4], person re-identification [5], [6] and robotics [7], [8], SOD extracts the most prominent objects in images to help scene analysis and understanding. Unlike semantic segmentation and instance segmentation [9], [10], SOD is insensitive to object categories, which means the foreground salient objects are category-agnostic [11]. In contrast to the inconspicuous background, the foreground is salient and contains a variety of different salient objects, such as cars, bicycles, cats, and dogs, *etc.*

Traditional SOD methods mainly use low-level features and certain priors, such as color contrast and background priors, to detect targets [12]. In recent years, CNN-based SOD methods [13], [14], [15], [16] have shown advantages over traditional hand-crafted feature-based methods in terms of model accuracy and generalization. The application of SOD is also extended from visible light images to multispectral ones [17]. RGB images are easily disturbed by the environment [18]. Thermal sensors rely on the thermal radiation of the object to generate images, which are not easily affected by variable conditions, such as weather, illumination, *etc*. [19], [20]. For example, the quality of thermal images is noticeably better than RGB images in low illumination. RGB-T image pairs have both the radiometric intensity of infrared and the detailed information of visible light. Compared with single RGB images, RGB-T multispectral fusion can generate discriminative and robust saliency features [21], [22]. Therefore, the RGB-T SOD method achieves a more robust generalization performance in real-world scenes.

To obtain accurate salient object results, many CNN-based models [23], [24], [25] focus on generating clear contours by learning edge maps. However, these methods ignore the relation learning between boundary pixels and foreground-background region pixels, resulting in unsatisfactory results. To tackle this issue, we propose a novel end-to-end position relation learning network (PRLNet). By integrating both distance relations and direction relations, PRLNet aims to enhance the inter-class separability of foreground and background features, as well as the intra-class compactness of salient object features.

The relative distance information between pixels can effectively alleviate the undesirable prediction of salient pixels [26].

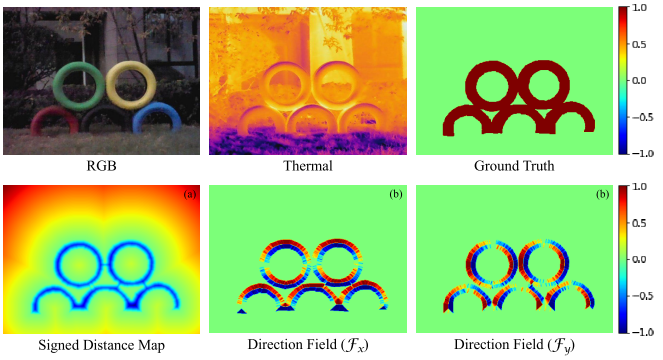


Fig. 1. RGB-T SOD with position-aware relation learning. **(a)** The signed distance map (SDM) calculates the distance from a pixel to the nearest boundary, where the sign indicates that the pixel is inside (+) or outside (-) the salient object. The zero level set is the boundary of the salient object. **(b)** The direction field ($\mathcal{F}_x, \mathcal{F}_y$) of a salient object points from the pixel to its nearest boundary pixel.

Inspired by the level set method [27], [28], the signed distance map (SDM) models the distance relation between regions and boundaries. As shown in Fig. 1 (a), SDM provides interaction information on boundaries based on level sets. Different from multi-task learning of SDM in decoder [29], [30], [31], we propose the SDM auxiliary module (SDMAM) to enhance the boundary-awareness of the encoder. SDMAM assists the encoder to learn the relative distance between the region pixels and the boundary, and increases the inter-class separation between foreground and background features.

Not only the distance relationship, but also the direction relationship between salient pixels is crucial in position-aware relationship learning. Fig. 1 (b) shows the visualization of the horizontal and vertical directions of the direction field [32]. As illustrated in Fig. 1 (b), the direction field can simply yet efficiently represent the directional information of salient intra-class pixels, which points from the nearest boundary pixel to the salient pixel. To strengthen the intra-class compactness of silent features, we propose a feature refinement approach with direction field (FRDF) to rectify the initial output feature maps of the decoder. Meanwhile, we design a novel direction-aware loss function to improve the smoothing loss [33], [34], which guides the model to generate homogeneous regions and sharp boundaries.

In this work, we use swin transformer as the backbone network. Finally, we propose a position-aware relation learning network (PRLNet) for RGB-T SOD. In summary, the main contributions of this paper are as follows.

- We propose a novel PRLNet to generate salient object masks with clear boundaries and homogeneous regions, which takes into account distance and direction relations. The proposed model consistently outperforms the state-of-the-art methods on three public RGB-T SOD datasets.
- Specifically, the SDM auxiliary module (SDMAM) is suggested to learn the distance relation of each pixel to the boundary, enhancing the inter-class separation of foreground features and background features.
- In order to strengthen the intra-class compactness, we design a feature refinement approach with direction field (FRDF) and direction-aware smoothness loss. The features close to the boundary are refined by utilizing the internal features of objects, reducing the intra-class variance of salient features.

The rest of this paper is organized as follows. Section II overviews the existing methods mainly on RGB and RGB-T SOD and swin transformer. In Section III, we introduce our proposed position-aware relation learning network for RGB-T SOD. Extensive experiments and visualization results on the three benchmark datasets are given in Section IV. Finally, we conclude our work in Section V.

II. Related Works

In this section, we review the previous SOD methods for RGB and RGB-T images. Meanwhile, related works about swin transformer are also included in this section.

A. RGB Salient Object Detection

Recently, most CNN-based SOD methods adopt a fully convolutional network (FCN) structure [35], [36]. To improve the accuracy of prediction results, multi-level feature fusion [37], [38], [39] and multi-task learning [23], [40] have been widely studied. Deng et al. [38] use the low-level and high-level features of FCN to learn residuals between intermediate saliency predictions and ground truth for refining saliency maps. Wu et al. [41] propose a cascaded partial decoder (CPD) that discards large-resolution features in shallow layers for acceleration, and fuses features in deep layers to obtain accurate saliency maps. Liu et al. [42] present pool-based modules to progressively refine features at multiple scales producing detailed results. The boundary prediction task [25] captures accurate boundary information of salient objects. Qin et al. [23] design a hybrid loss for predicting the boundaries of salient objects. However, boundary supervision lacks consideration of the interaction between boundary pixels and target pixels. Inspired by the level set method [27], [28], we develop a novel signed distance map auxiliary module (SDMAM) to improve encoder features. SDMAM takes into account the distance relation of pixels in boundary neighborhoods. The distance relationship between foreground-background region pixels and boundary pixels can effectively enhance the inter-class separability of features.

B. RGB-T Salient Object Detection

Compared to RGB images, multimodal data offer more information on salient objects [43]. In recent years, synergistic SOD between thermal and visible images has been widely studied [44], [45], [46]. The dual encoders extract RGB-T features respectively, and the decoder outputs the salient prediction results [47]. The RGB-T SOD methods take full advantage of the complementary capabilities between multimodal sensors to generate cross-modal robust fusion features [48], [49], [50]. Tu et al. [51] suggest a collaborative graph learning algorithm that uses superpixels as graph nodes to learn RGB-T node saliency. Zhang et al. [13] transform multi-spectral SOD into a CNN feature fusion problem, and propose to capture semantic information and visual details of RGB-T at different depths by fusing multi-level CNN features. Tu et al. [52]

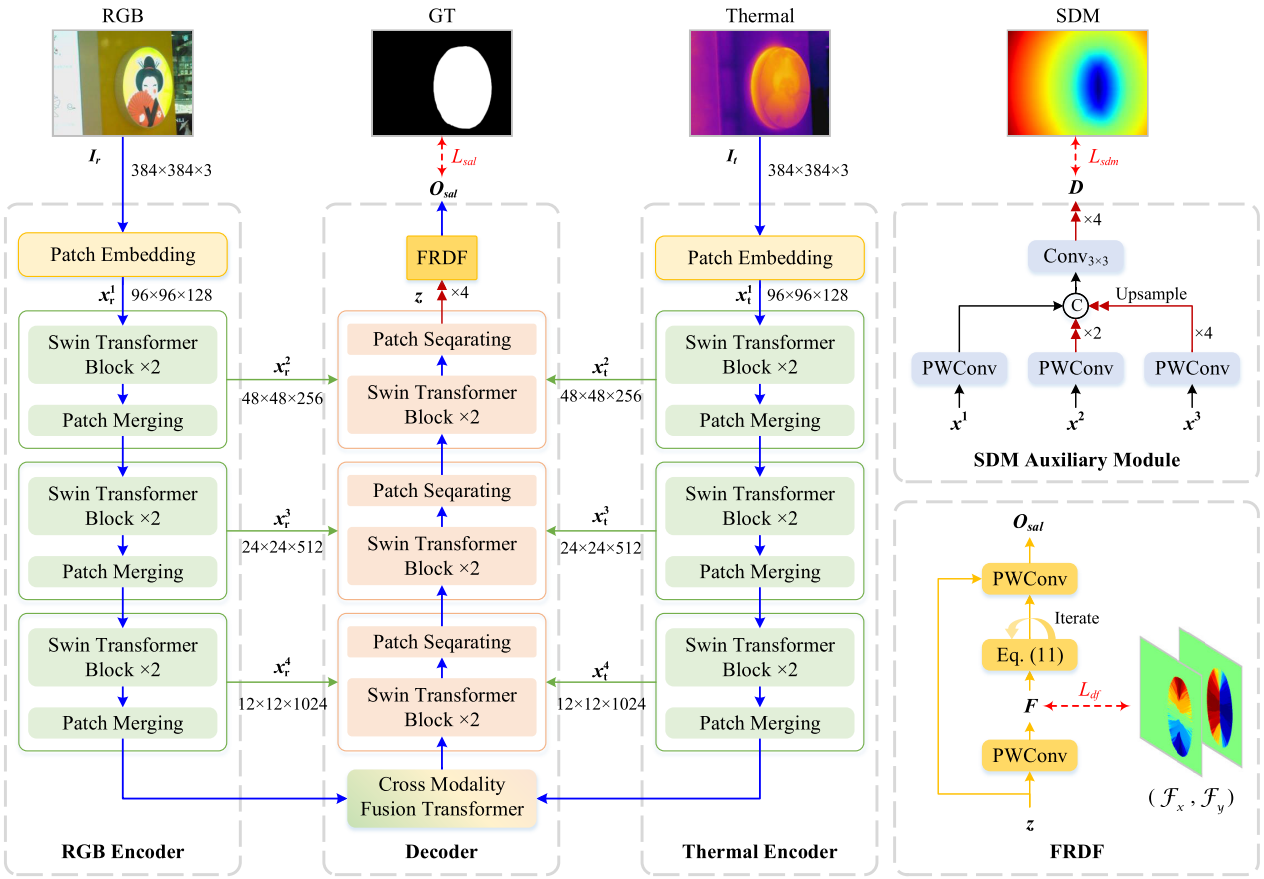


Fig. 2. The framework of our proposed PRLNet. Our network consists of four main parts, namely dual-stream encoders, RGB-T decoder, SDM auxiliary module and a feature refinement approach with direction fields (FRDF). First, multiscale features of RGB-T are extracted by a dual-stream swin transformer encoder. (Sec. III-A). Then, we construct SDMAM for encoders to learn the distance relationship between regional pixels and boundary pixels (Sec. III-B). Next, the reverse swin transformer decoder aggregates the complementarity between different levels of RGB-T features (Sec. III-C), where © denotes concatenation. In addition, FRDF is further designed by exploring direction information between salient pixels to strengthen the intra-class compactness of the salient features (Sec. III-D). Finally, we propose a novel position-aware relation learning loss to generate object masks with clear boundaries and homogeneous regions (Sec. III-E).

exploit the complementarity of different modalities of image content and multiple types of cues to extract multi-level multimodal features. Zhou et al. [53] propose an effective and consistent feature fusion network that combines features of different levels through a multi-level consistent fusion module to obtain complementary information. In this paper, to handle long-range dependencies between RGB-T, we develop a cross-spectra fusion transformer. Furthermore, we propose a feature refinement approach with direction field (FRDF) to enhance the intra-class compactness of salient objects. FRDF exploits feature far from the boundary to refine the features of pixels close to the boundary.

C. Swin Transformer

Compared with CNN, transformer has an advantage in modeling long-range dependencies [54], [55]. ViT [56] and DETR [57] apply transformer to computer vision tasks and achieve promising performance. However, the computational complexity of the transformer is proportional to the square of the image size. To handle high-resolution images, swin transformer [58] introduces the hierarchical structure commonly used in CNN and achieves SOTA results on dense prediction tasks. Swin transformer gradually becomes a powerful general backbone network for SOD [59]. Liu et al. [60] propose a cross-modal fusion network based on the swin transformer for RGB-T SOD, bridging the gap between two modalities through an attention mechanism. Zhu et al. [61] encode multi-scale features via the swin transformer in a coarse-to-fine manner to learn salient region feature representations. In this paper, swin transformer block is used as the backbone for both the encoder stage and decoder stage. Specifically, we employ dual-swin transformer encoders to extract multi-scale features from RGB and thermal images, respectively. Referring to the patch merging layer, we design a patch separating layer to decode RGB-T hierarchical features and generate robust results with multispectral complementarity.

III. PRLNet

In this section, we elaborate on PRLNet for RGB-T SOD with swin transformer. The overall architecture is illustrated in Fig. 2, which consists of four main parts: Dual-stream encoders for both RGB-T images, a decoder for pixel-by-pixel prediction, an SDM auxiliary module (SDMAM) and a feature refinement approach with direction fields (FRDF). They are simultaneously optimized during the training process.

As shown in Fig. 2, PRLNet takes the RGB-T image pair as input, and segments the precise mask of the salient objects. We first use the dual-stream swin transformer encoder to generate multi-scale features of RGB and thermal images (Sec. III-A). Then, to improve the boundary perception of the encoder, we introduce SDMAM to learn the distance relationship between regional pixels and boundary pixels. SDMAM enhances the separability of foreground-background features (Sec. III-B). Next, we design a patch separating layer and construct an inverse swin transformer, which aggregates different levels of RGB-T features (Sec. III-C). To facilitate the robust cross-spectral features from the decoder, we further refine them with the direction information between salient pixels to strengthen the intra-class compactness of the feature for different salient objects (Sec. III-D). Finally, benefiting from the effective learning of position relations, we present a position-aware relation learning loss function to generate object masks with clear boundaries and homogeneous regions (Sec. III-E). The pipeline of PRLNet is illustrated in Algorithm 1.

A. Dual-Stream Swin Transformer Encoder

Swin Transformer introduces hierarchical feature mapping and shifted window attention, which has both the advantages of transformer and CNN structure [58]. We employ two swin transformers to extract efficient features for RGB-T image pairs, respectively. Concretely, the images are first divided into 4×4 patches and then input to the patch embedding layer, which is a 4×4 convolution with stride 4. Next, as shown in Fig. 2, RGB-T salient features are extracted by three swin transformer layers (ST), consisting of swin transformer block (STB) and patch merging layer (PM). That is,

$$\begin{aligned} \mathbf{R} &= \{\mathbf{x}_r^i\}_{i=1}^4 = \mathrm{ST}_r\left(\boldsymbol{I}_r\right), \\ \mathbf{T} &= \{\mathbf{x}_t^i\}_{i=1}^4 = \mathrm{ST}_t\left(\boldsymbol{I}_t\right). \end{aligned} \tag{1}$$

The dual encoder outputs the hierarchical representation $\mathbf{R}$ and $\mathbf{T}$, where $\mathbf{x}_r^i$ and $\mathbf{x}_t^i$ denote the i-th layer features of the RGB and thermal encoder, respectively. $\boldsymbol{I}_r$ and $\boldsymbol{I}_t$ indicate the RGB and thermal images, which are the input of the encoder. The bold symbols indicate the matrix. The ST_r and ST_t functions are composed of STB and PM, and represent the standard swin transformer backbones in RGB and thermal branches, respectively.

Different from ViT block [56], STB replaces the multi-head attention mechanism (MSA) of ViT with window-based MSA (W-MSA) and shifted window-based MSA (SW-MSA). More formally, STB is defended as

$$\begin{aligned} \hat{z}^l &= \text{W-MSA}\left(\mathrm{LN}\left(z^{l-1}\right)\right) + z^{l-1}, \\ z^l &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l, \\ \hat{z}^{l+1} &= \text{SW-MSA}\left(\mathrm{LN}\left(z^l\right)\right) + z^l, \\ z^{l+1} &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1}, \end{aligned} \tag{2}$$

where $\hat{z}^l$ and z^l denote the output feature of the (S)W-MSA module and the MLP module for block l, respectively. Fig. 3 demonstrates the detailed architecture of STB. STB uses a layernorm (LN) layer before each MSA module and each multilayer perceptron (MLP), followed by residual connections. As illustrated in Fig. 4 (a), PM reduces the resolution of the features and increases the number of channels of the features.

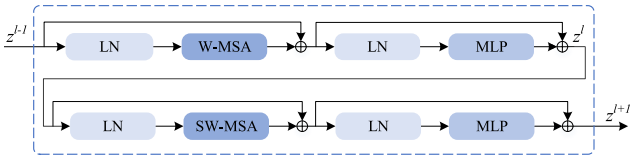
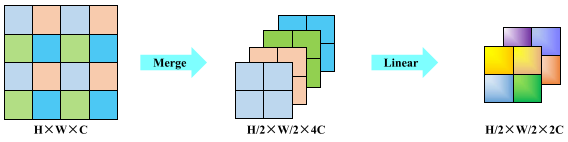
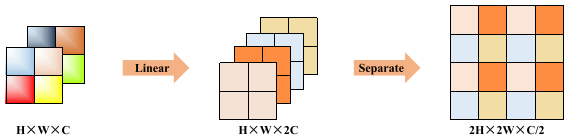


Fig. 3. The architecture of the swin transformer block (STB). W-MSA calculates the pairwise attention of each token in the window. SW-MSA shifts the window of W-MSA by half the window length.



(a) Patch Merging



(b) Patch Separating

Fig. 4. **(a)** Patch merging layer (PM) merges the neighboring patches into a new patch, thus reducing the resolution. **(b)** Our proposed patch separating layer (PS) upsamples features by expanding each patch into multiple sub-patches.

B. SDM Auxiliary Module

The signed distance map (SDM) [27], [28] models the distance relationship between pixels in the foreground-background region and the boundary, and further distinguishes foreground and background with positive and negative signs. According to the ground truth $\boldsymbol{G}$, SDM transformation $\mathcal{D}(p)$ for each pixel $p \in \boldsymbol{G}$ is given by:

$$\mathcal{D}(p) = \begin{cases} -\inf\limits_{\forall b \in \partial\mathcal{S}} d(p,b), & p \in \mathcal{S}_{\text{sal}} \\ 0, & p \in \partial\mathcal{S} \\ +\inf\limits_{\forall b \in \partial\mathcal{S}} d(p,b), & p \in \mathcal{S}_{\text{bg}} \end{cases} \tag{3}$$

where inf denotes the infimum, b is the boundary pixel. In Eq. (3), $\partial\mathcal{S}$ is the zero level set which represents the pixel set of the target boundary. $\mathcal{S}_{\text{sal}}$ and $\mathcal{S}_{\text{bg}}$ indicate the salient object pixel set and background pixel set, respectively. In our work, $d(\cdot)$ indicates the Euclidean distance. As shown in Fig. 1 (a), SDM not only perceives the boundary of an object, but also predicts whether the pixel is located inside or outside the object. For each pixel $p \in \boldsymbol{G}$, the sign of $\mathcal{D}(p)$ indicates whether it is located outside (i.e., $\mathcal{D}(p) > 0$) or inside (i.e., $\mathcal{D}(p) < 0$) the object. $\mathcal{D}(p) = 0$ denotes the boundary of the object. $|\mathcal{D}(p)|$ represents the distance from pixel p to the boundary.

To precisely perceive the boundaries of salient objects, we present an SDM auxiliary module (SDMAM) to learn the

distance relation between region pixels and boundary pixels. Benefiting from SDM, SDMAM can effectively strengthen the inter-class separability of foreground-background region features. The upper right part of Fig. 2 shows the structure of SDMAM in detail. The shallow high-resolution features contain rich texture information. SDMAM integrates RGB-T shallow features to predict the distance relationship between pixels. Formally,

$$\boldsymbol{D} = \text{SDMAM}\left(\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\right), \tag{4}$$

where $\mathbf{x}^i = \text{concat}\left(\mathbf{x}_r^i, \mathbf{x}_t^i\right)$, $i = 1, 2, 3$. $\boldsymbol{D} \in \mathbb{R}^{h\times w\times 1}$ represents the prediction result of SDMAM. The dimensions of $\mathbf{x}^1$, $\mathbf{x}^2$ and $\mathbf{x}^3$ are $\mathbb{R}^{\frac{h}{4}\times\frac{w}{4}\times c}$, $\mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times 2c}$ and $\mathbb{R}^{\frac{h}{16}\times\frac{w}{16}\times 4c}$, respectively. In this paper, $h = w = 384$, $c = 128$.

Specifically, the multi-scale features $\{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\}$ are further fused by pointwise convolution (PWConv) [62] with ReLU and upsampling operations,

$$\mathbf{y}^i = \left[\text{PWConv}\left(\mathbf{x}^i\right)\right]^{\times(2^{i-1})}, \tag{5}$$

where $i = 1, 2, 3$. $[\cdot]^{\times(n)}$ denotes upsampling the features by n times. In Eq. (5), the different scale high-resolution features $\mathbf{y}^i \in \mathbb{R}^{\frac{h}{4}\times\frac{w}{4}\times 32}$. Finally, the multi-scale multi-spectral features $\mathbf{y}$ are fused by 3×3 convolution and upsampled to the resolution of the input image.

$$\begin{aligned}\mathbf{y} &= \text{concat}\left(\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3\right),\\ \boldsymbol{D} &= \tanh\left[\text{Conv}_{3\times 3}(\mathbf{y})\right]^{\times(4)},\end{aligned} \tag{6}$$

where the output of SDMAM is $\boldsymbol{D} \in \mathbb{R}^{h\times w\times 1}$, $\text{Conv}_{3\times 3}$ indicate 3×3 convolution with stride 1.

SDMAM outputs boundary-aware features, which contain distance relations between pixels of different classes, enhancing inter-class separability between salient foreground objects and background.

C. Reverse Swin Transformer Decoder

Our decoder is designed to decode patches as saliency maps. Hence, we propose a novel patch upsampling method with multi-level patch fusion and a patch-based SOD decoder.

1) Cross Spectrum Fusion Transformer: Concretely, the RGB-T encoder feature map $\mathbf{x}^4 = \text{concat}\left(\mathbf{x}_r^4, \mathbf{x}_t^4\right)$ is flattened into an input sequence. A set of queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ is computed by embedding the input sequence into three weight matrices.

In Eq. (7), we compute cross-spectral attention $\mathbf{z}^4$ as in [22] and [55].

$$\mathbf{z}^4 = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \tag{7}$$

where $\mathbf{z}^4 \in \mathbb{R}^{\frac{h}{32}\times\frac{w}{32}\times 8c}$, $\sqrt{d}$ is an adjustment factor that prevents the softmax function from having too large an input value resulting in too small a partial derivative.

2) Reverse Swin Transformer Decoder: In swin transformer, the patch merging (PM) integrates patches of different windows to reduce the spatial resolution of feature maps. Inspired by PM, we design patch separating (PS) to upsample patches by separating each patch for multiple sub-patches. As shown in Fig. 4 (b), Based on PS, we propose a reverse swin transformer layer (RST) for the decoder.

The reverse swin transformer decoder is illustrated in the middle of Fig. 2. RST layer includes STB and PS. For RGB-T features of encoders, RST generates more patches and progressively decodes the patches into high-resolution saliency maps, as in Eq. (8).

$$\mathbf{z}^i = \text{RST}\left(\mathbf{z}^{i+1}, \mathbf{x}^{i+1}\right), \tag{8}$$

where $i = 1, 2, 3$. The dimensions of $\mathbf{z}^1$, $\mathbf{z}^2$ and $\mathbf{z}^3$ are $\mathbb{R}^{\frac{h}{4}\times\frac{w}{4}\times c}$, $\mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times 2c}$ and $\mathbb{R}^{\frac{h}{16}\times\frac{w}{16}\times 4c}$, respectively. The salient swin transformer decoder output $\mathbf{z} \in \mathbb{R}^{h\times w\times 64}$ is obtained by upsampling $\mathbf{z}^1$ by a factor of 4.

D. Feature Refinement Approach With Direction Field

The direction field (DF) [63] offers the direction relationship between salient pixels. The direction vector of each pixel points from the boundary to the center. The mathematical definition of the direction field function $\mathcal{F}$ is shown in Eq. (9). The direction of $\mathcal{F}(p)$ is from b pointing to p, and b is the nearest pixel to p on the boundary. For the pixel $p \in \boldsymbol{G}$,

$$\mathcal{F}(p) = \begin{cases} \inf\limits_{\forall b\in\partial\mathcal{S}} \vec{bp}, & p \in \mathcal{S}_{\text{sal}} \\ (0, 0), & p \in \mathcal{S}_{\text{bg}} \end{cases} \tag{9}$$

where $\mathcal{S}_{\text{sal}}$ and $\mathcal{S}_{\text{bg}}$ denote the salient object pixel set and background pixel set, respectively.

The refinement of the initially predicted features provides an effective way to improve salient object masks. Based on this idea, we design a feature refinement approach with direction field (FRDF). With the help of directional information, FRDF uses features inside the object to improve the visual representation near the boundary. FRDF progressively enforces the intra-class compactness of salient region features through several iterative updates. As shown in the bottom right of Fig. 2, we first use the decoder feature $\mathbf{z}$ to predict the direction field feature $\boldsymbol{F} \in \mathbb{R}^{h\times w\times 2}$ in Eq. (10).

$$\boldsymbol{F} = \text{PWConv}\left(\mathbf{z}\right). \tag{10}$$

Then, the initial predicted saliency feature map is refined step by step iteratively according to Eq. (11).

$$\mathbf{z}_k(p) = \mathbf{z}_k\left(p_x + \boldsymbol{F}_x(p), p_y + \boldsymbol{F}_y(p)\right), \tag{11}$$

where $\mathbf{z}_k$ denotes the salient feature map after the k-th iteration. The number of iterations is set as $K = 5$, which is further ablated with experiments in Sec. IV-D.3. p_x and p_y indicate the x and y coordinates of pixel p, respectively. The output of FRDF is the refined feature $\mathbf{z}^* \in \mathbb{R}^{h\times w\times 2c}$. FRDF efficiently exploits direction priors to reduce the intra-class variability of salient features in a supervised manner.

Finally, the PWConv layer combines initial feature $\mathbf{z}$ with the rectified feature $\mathbf{z}^*$ to generate the salient mask

$\boldsymbol{O}_{sal} \in \mathbb{R}^{h\times w\times 1}$. Both SDM and DF are learned in a fully-supervised way, which will be discussed in Sec. III-E. Based on the ground truth $\boldsymbol{G}$, we obtain the true supervised signal for the SDMAM and FRDF.

E. Loss Function

According to the ground truth $\boldsymbol{G}$ of the image, the true SDM and the true direction field of the salient object can be calculated by mathematical models, *i.e.*, Eq. (3) and Eq. (9).

$$\begin{aligned} \boldsymbol{D}_{gt} &= \mathcal{D}(\boldsymbol{G}), \\ \boldsymbol{F}_{gt} &= \mathcal{F}(\boldsymbol{G}), \end{aligned} \tag{12}$$

In Eq. (12), $\boldsymbol{D}_{gt}$ is the true SDM and $\boldsymbol{F}_{gt}$ is the true DF. They guide SDMAM and FRDF to enhance intra-class compactness and inter-class separability, which are weakly supervision for RGB-T SOD.

1) SDM Loss: SDM loss is

$$\mathcal{L}_{sdm} = \sum_{p\in\Omega} \left\| \boldsymbol{D} - \boldsymbol{D}_{gt} \right\|^2, \tag{13}$$

where Ω denotes all pixels, $\boldsymbol{D}$ is the predicted result of SDMAM. $\mathcal{L}_{sdm}$ drives PRLNet to learn the distance relationship between foreground-background regions and boundaries, effectively enhancing the inter-class differences of salient features.

2) Direction Field Loss: DF loss is

$$\mathcal{L}_{df} = \sum_{p\in\Omega} \left(\|\boldsymbol{F} - \boldsymbol{F}_{gt}\|_2 + \left\| \cos^{-1}\langle \boldsymbol{F}, \boldsymbol{F}_{gt} \rangle \right\|^2 \right), \tag{14}$$

where $\boldsymbol{F}$ and $\boldsymbol{F}_{gt}$ indicate the predicted DF and the corresponding ground truth, respectively. $\mathcal{L}_{df}$ guides the model to learn the direction relationship between pixels, which rectifies features of boundary neighborhood by exploiting the features inside salient objects.

3) Direction-Aware Smoothness Loss: We develop a novel direction-aware smoothness loss ($\mathcal{L}_{DS}$) that enhances the compactness of regions and the boundary clearness. We calculate the first-order derivative of the saliency map in the smooth term [33], [52]. $\mathcal{L}_{DS}$ is defined as follows,

$$\mathcal{L}_{DS}(\boldsymbol{O}, \boldsymbol{G}) = \sum_{p\in\Omega} \sum_{\partial_{x,y}} w\left(p\right) \psi \left(\left| \partial \boldsymbol{O} \right| e^{-\alpha |\partial \boldsymbol{G}|} \right), \tag{15}$$

$$w(p) = \begin{cases} \|\mathcal{F}(p)\|^{-1}, & p \in \mathcal{S}_{\text{sal}} \\ 1, & p \in \mathcal{S}_{\text{bg}} \end{cases} \tag{16}$$

where $\psi(m) = \sqrt{m^2 + 0.001^2}$, $\boldsymbol{O}$ and $\boldsymbol{G}$ represent the predicted salient result and ground truth, respectively. $\partial_{x,y}$ denotes the partial derivatives in x and y directions. In Eq. (15), same as [34], we set $\alpha = 10$ to control the weight of the boundary. In Eq. (16), $w\left(p\right)$ indicates the weight on pixel p. Therefore, the saliency loss is

$$\mathcal{L}_{sal} = \mathcal{L}_{DS}\left(\boldsymbol{O}_{sal}, \boldsymbol{G}\right). \tag{17}$$

Finally, our position-aware relation learning loss $\mathcal{L}_{prl}$ is

$$\mathcal{L}_{prl} = \mathcal{L}_{sal} + \lambda_1 \mathcal{L}_{sdm} + \lambda_2 \mathcal{L}_{df}, \tag{18}$$

Algorithm 1 The Pipeline of PRLNet

Input: RGB-T images $\{\boldsymbol{I}_r, \boldsymbol{I}_t\}$, ground truth $\boldsymbol{G}$
Output: Salient object mask $\boldsymbol{O}_{sal}$, signed distance map $\boldsymbol{D}$, direction field $\boldsymbol{F}$

```
Init true SDM D_gt ← 𝒟(G) using Eq. (3)
Init true direction field F_gt ← ℱ(G) using Eq. (9)
Init the number of FRDF iterations K = 5
while epoch < N do
    Extract RGB-T features R and T using Eq. (1)
    Generate signed distance map D using Eq. (4)
    Generate decoder output feature z using Eq. (8)
    Generate direction field F using Eq. (10)
    for k ≤ K do
        Iterative refinement of decoder feature
        z ← z + F using Eq. (11)
    end for
    Generate salient mask result O_sal using initial
    features z and refined features z*
    Calculate the SDM loss
    L_sdm ← ℒ(D, D_gt) using Eq. (13)
    Calculate the DF loss
    L_df ← ℒ(F, F_gt) using Eq. (14)
    Calculate the direction-aware smoothness loss
    L_sal ← ℒ(O_sal, G) using Eq. (17)
    Calculate the overall loss function of PRLNet
    L_prl ← L_sal + λ1 L_sdm + λ2 L_df,
    O_sal, D, F ← arg min L_prl
end while
return O_sal
```

where λ_1 and λ_2 are the hyper-parameters controlling the contributions of the two losses, which are set via ablative analysis in Sec. IV-D.3. The proposed PRLNet is optimized through Eq. (18) jointly. Our proposed PRL loss can effectively guide the network to pay more attention to the pixels around the object boundary, thereby helping the network to predict salient masks with sharp boundaries and homogeneous regions.

IV. Experiments

In this section, we first introduce the three RGB-T SOD datasets, implementation details, and evaluation metrics. We then give the details of our experiments. In particular, we evaluate our method on three widely used datasets to compare with SOTA methods. Moreover, ablation studies are also conducted to further validate the validity of our network.

A. Experimental Setup

1) Datasets: There are three available benchmark datasets for RGB-T SOD tasks, including VT821 [44], VT1000 [51] and VT5000 [47], which have 821, 1000, and 5000 aligned image pairs, respectively. Compared with VT821, the VT1000 dataset has more images and scenes, and the quality of the thermal images is better. VT5000 provides a large-scale dataset for RGB-T SOD. In addition, VT5000 does not require manual RGB-T image pair alignment, which reduces the errors caused by manual alignment. VT5000 contains a variety of

TABLE I
DETAILS OF 13 CHALLENGES IN VT5000 DATASET

Challenge	Describe
BSO	Big salient object: the proportion of pixels of salient objects to the image is more than 0.26.
SSO	Small salient object: the percentage of the number of salient pixels is less than 0.05.
MSO	Multiple salient objects.
CB	Center bias: the salient object is out of the center of the image.
CIB	Cross image boundary: a part of the salient object is outside the image.
OF	Out of focus: out of focus causes the whole image to be blurred.
SA	Similar appearance: the salient object is similar to the color and texture of the background.
TC	Thermal crossover: the salient object is similar to its surrounding temperature.
IC	Image clutter: the scene is cluttered.
LI	Low illumination: the scene is cloudy or at night.
BW	Bad weather: the scene is rainy or foggy.
RGB	Objects are not clear in RGB images.
T	Objects are not clear in the thermal image.

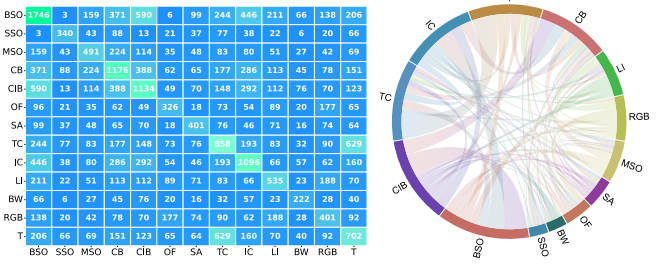


Fig. 5. **Left**: Co-challenges distribution over VT5000 datasets. The number in each grid indicates the total number of images. **Right**: Multi-dependencies among these challenges. A larger arc length indicates a higher probability of one challenge correlating to another.

complex scenes with diverse objects and covers 13 challenges of RGB-T SOD [47]. Challenge descriptions are provided in TABLE I, and the co-challenge distribution is shown in Fig. 5. In TABLE I, VT5000 simulates image saliency detection under real-world conditions mainly in terms of target diversity (BSO, SSO, MSO, CB, CIB, OF, SA and TC), scene complexity (IC, LI and BW) and spectral effectiveness (RGB and T).

2) Implementation Details: To extract multispectral features, we initialize our backbone networks through the parameters of the pre-trained Swin-B model [58]. The whole network is then trained on a large-scale dataset with the proposed position-aware relation learning loss in an end-to-end manner.

For a fair comparison, we use the same setting as in [38], [51], [60], and [47] where half of the VT5000 dataset is applied as the training set. VT821, VT1000, and the other half of VT5000 are treated as the test set. In addition, each image is random flipping, cropping and rotation ($-15^\circ \sim 15^\circ$), and then resized to 384×384 . We train our models by using the adaptive optimizer Adam. The initial learning rate of the network is set to 10^{-5} and is decayed by 0.1 every 100 epochs. The total epoch number is set to 300. The mini-batch size is set as 6. Our framework is implemented by PyTorch. The experiment is conducted on a computer with 3.0 GHz CPU, 128 GB RAM, and four NVIDIA GeForce RTX 3090 GPUs.

B. Evaluation Metrics

To facilitate the comparison of the performance of different RGB-T methods, we use the evaluation metrics commonly used in the SOD model: P-R curves [64], S-measure ($S_\alpha \uparrow$) [65], F-measure ($F_\beta \uparrow$) [66], E-measure ($E_m \uparrow$) [67] and MAE ($\mathcal{M} \downarrow$) [68]. $\uparrow$ and $\downarrow$ indicate that the higher the better and the lower the better, respectively. The P-R curves and F_β evaluate the quality of the prediction results in terms of Precision and Recall. S_α and E_m mainly measure the structural similarity between the predicted saliency mask and GT. $\mathcal{M}$ counts the error of the incorrectly predicted pixels. We use the above metrics to evaluate the model accurately and comprehensively. The formal definition is as follows.

1) P-R Curves: We first demonstrate the performance of our model through standard P-R curves [64]. Different thresholds ([0, 255]) are applied to the prediction to generate a binarized result that produces pairs of Precision-Recall values. A set of thresholds provides the P-R curve of the model. Formally, the P and R are defined based on the binarized salient object mask and the corresponding ground truth in Eq. (19).

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{19}$$

where TP, FP and FN denote true positive, false positive and false negative, respectively.

2) S-Measure: The structure measure (S_α) can effectively evaluate the spatial structure compactness between prediction and ground truth [65].

$$\mathrm{S}_\alpha = \alpha \mathrm{S}_o + (1 - \alpha)\mathrm{S}_r, \tag{20}$$

where α is set as 0.5 empirically [65]. In Eq. (20), S_α integrates object-aware structural similarity S_o and region-aware structural similarity S_r.

3) F-Measure: F_β takes into account precision and recall [66], and calculates the weighted harmonic mean of P and R:

$$\mathrm{F}_\beta = \frac{\left(1 + \beta^2\right) \times \mathrm{P} \times \mathrm{R}}{\beta^2 \times \mathrm{P} + \mathrm{R}}, \tag{21}$$

where we set $\beta^2 = 0.3$ to weigh precision more than recall.

4) E-Measure: The enhanced-alignment measure metric (E_m) considers both local pixel values and image-level averages. E_m captures image-level statistics and local pixel matching information [67].

$$\mathrm{E}_m = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \phi_{ij}. \tag{22}$$

In Eq. (22), H and W are the height and width of the object map, respectively. ϕ is the enhanced alignment matrix [67].

5) MAE: The mean absolute error ($\mathcal{M}$) [68] measures the difference between saliency prediction $\boldsymbol{O} \in [0, 1]^{H \times W}$ and ground truth mask $\boldsymbol{G} \in \{0, 1\}^{H \times W}$,

$$\mathcal{M} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} |\boldsymbol{O}_{ij} - \boldsymbol{G}_{ij}|. \tag{23}$$

TABLE II

QUANTITATIVE COMPARISON WITH SOTA METHOD ON THREE BENCHMARK DATASETS IN TERMS OFS-MEASURE ($S_\alpha \uparrow$), F-MEASURE ($F_\beta \uparrow$), E-MEASURE ($E_m \uparrow$) ANDMAE ($\mathcal{M} \downarrow$).↑ AND ↓ REPRESENT THE HIGHER THE BETTER AND THE LOWER THE BETTER, RESPECTIVELY. THE BEST RESULT IN EACH COLUMN IS IN **RED**, AND THE SECOND IS IN BLUE

Methods	VT821				VT1000				VT5000			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$
R3Net [38]	0.785	0.809	0.660	0.073	0.842	0.859	0.761	0.055	0.757	0.790	0.615	0.083
CPD [41]	0.818	0.718	0.843	0.079	0.907	0.863	0.923	0.031	0.855	0.787	0.894	0.046
PoolNet [42]	0.751	0.739	0.578	0.109	0.834	0.813	0.714	0.067	0.769	0.755	0.588	0.089
SGDL [51]	0.765	0.847	0.731	0.085	0.787	0.856	0.764	0.090	0.750	0.824	0.672	0.089
ADF [47]	0.810	0.716	0.842	0.077	0.910	0.847	0.921	0.034	0.863	0.778	0.891	0.048
FMCF [13]	0.760	0.796	0.640	0.080	0.873	0.899	0.823	0.037	0.814	0.864	0.734	0.055
MIDD [52]	0.871	0.804	0.895	0.045	0.915	0.882	0.933	0.027	0.867	0.801	0.897	0.043
ECFFNet [53]	0.877	0.810	0.902	0.034	0.923	0.876	0.930	0.021	0.874	0.806	0.906	0.038
SwinNet [60]	0.904	0.847	0.926	0.030	0.938	0.896	0.947	0.018	0.912	0.865	0.942	0.026
PRLNet (Ours)	**0.917**	**0.860**	**0.932**	**0.025**	**0.944**	**0.902**	**0.951**	**0.016**	**0.921**	**0.875**	**0.948**	**0.023**

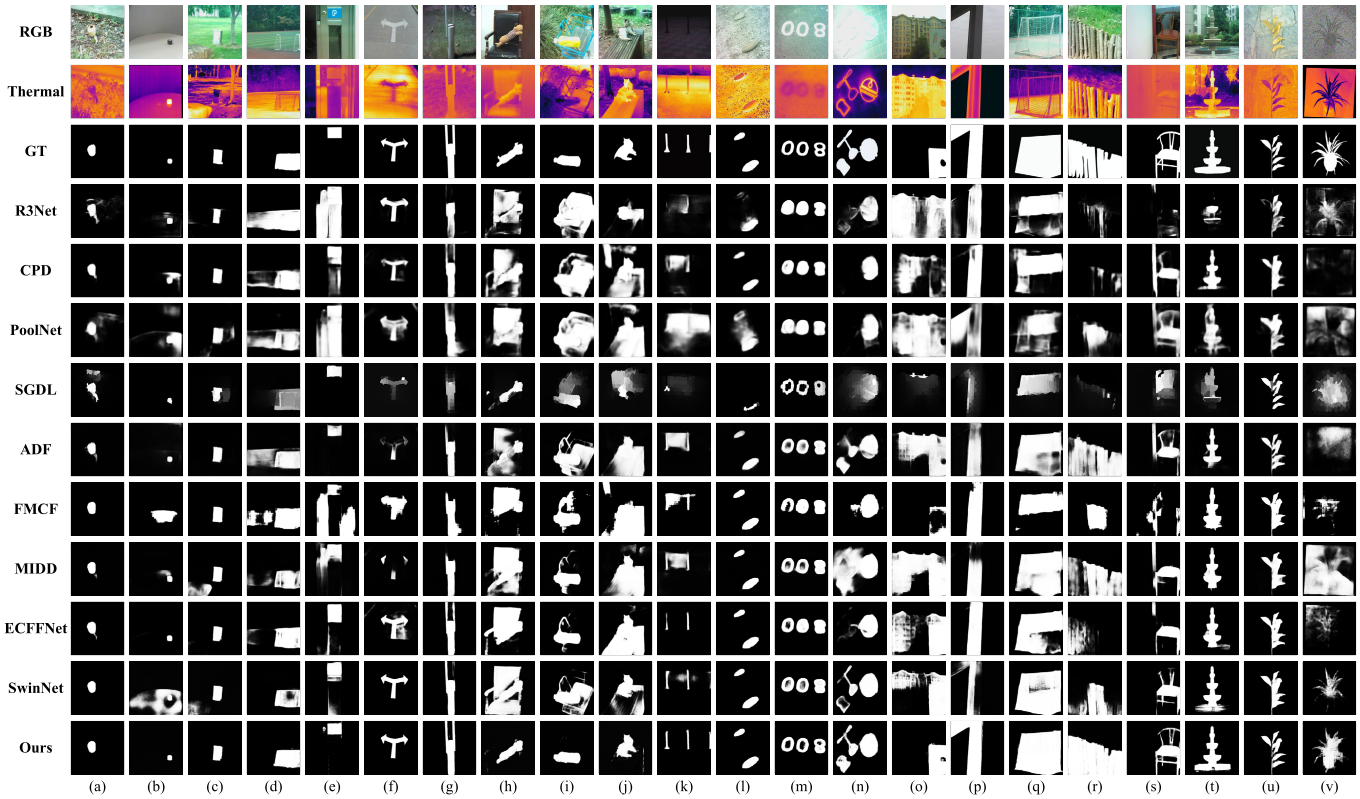


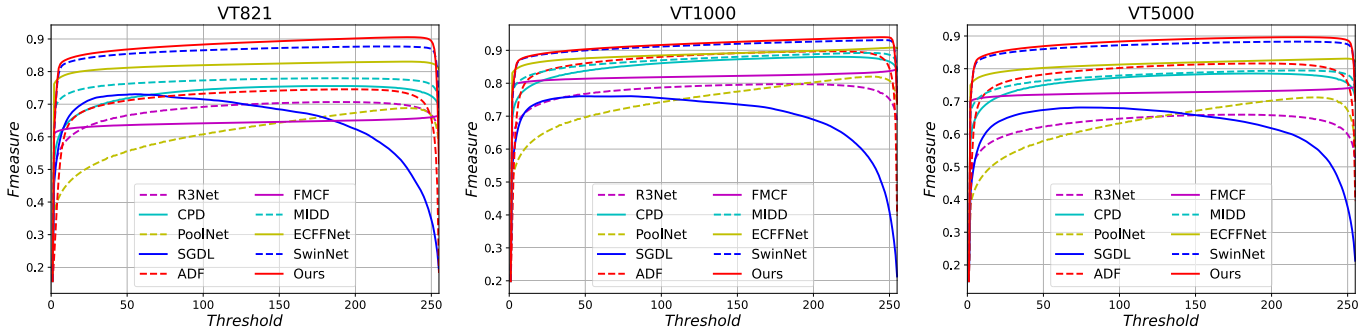
Fig. 6. Visual comparisons of different SOTA methods under various challenges, where each column indicates one input image. This figure shows that our proposed method (Ours) consistently generates saliency maps close to the Ground Truth (GT). Zoom in for details.

C. Comparison With State-of-the-Art Methods

To evaluate the validity of the proposed PRLNet, we conduct experiments compared with state-of-the-art methods on three datasets, which are shown in TABLE II and Fig. 6, 7, 8. Three RGB SOD methods include R3Net [38], CPD [41] and PoolNet [42]. Six RGB-T SOD methods include SGDL [51], ADF [47], FMCF [13], MIDD [52], ECFFNet [53] and SwinNet [60].

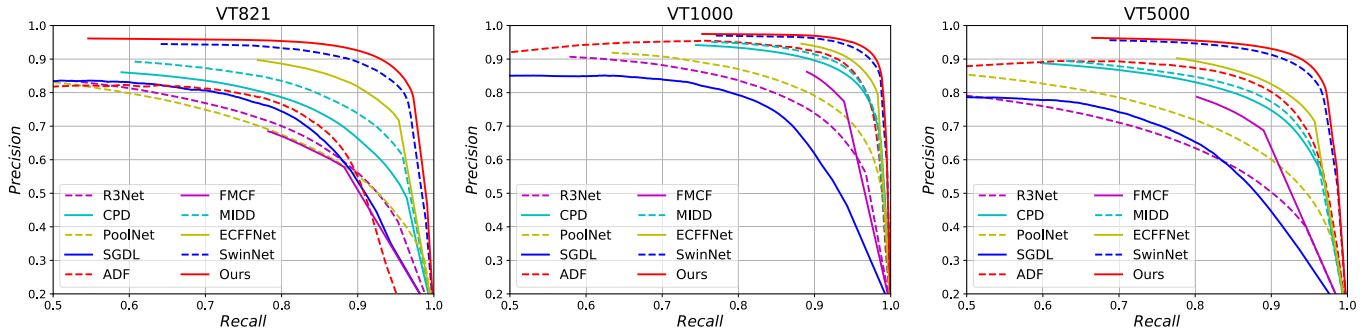
1) Qualitative Comparison: The results visualized in Fig. 6 display a qualitative comparison of some challenging image pairs, such as SSO (column (a)-(c)), CB (column (d) and (e)), BSO (column (e), (f) and (q)-(t)), BW (column (g) and (r)), TC (column (h)-(j)), LI (column (f) and (k)), MSO (column (k)-(n)), SA (column (i)-(l)), CIB (column (o)-(r) and (u)), IC (column (s) and (v)), OF (column (p) and (r)), RGB images with low quality (columns (a), (g), (k), (n) and (v)) and thermal images with low quality (columns (a), (e), (i), (m) and (s)). As illustrated in Fig. 6, the results of our PRLNet are qualitatively superior to all SOTA methods. Our method takes full advantage of the discriminative feature representation capabilities, while taking the position relations between pixels into account, *i.e.*, distance and direction relationships.

As shown in Fig. 6 (e), (h) and (o), the salient objects and background objects in certain spectral images have similar intensities, which can lead to confusion between



(a) VT821 dataset (b) VT1000 dataset (c) VT5000 dataset

Fig. 7. F-measure curves comparison of different methods on VT821, VT1000 and VT5000 datasets. The F-measure curves show that the salient maps of our PRLNet (Ours) usually are closer to the true object mask.



(a) VT821 dataset (b) VT1000 dataset (c) VT5000 dataset

Fig. 8. P-R curves comparison of different methods on VT821, VT1000 and VT5000 datasets. The P-R curves show that our PRLNet (Ours) consistently outperforms the SOTA models on three datasets.

foreground and background classes. Our proposed SDMAM effectively addresses this problem by explicitly constraining the foreground-background difference with signs and modeling the distance of different pixels from the boundary. SDMAM increases inter-class separability. From the results in Fig. 6 (e), (h) and (o), it can be observed that for background objects similar to the target, such as *cabinets*, *chairs* and *buildings*, our method accurately excludes inter-class interference.

On the other hand, the foreground objects also contain many components with large differences, which leads to inconsistencies in the intra-class features, as observed in Fig. 6 (d), (n) and (q). Our proposed FRDF learns the directional relations of pixels in salient regions, enhancing the intra-class compactness of feature representations. From the results in Fig. 6 (d), (n) and (q), it can be observed that the salient object masks generated by our model are more homogenous compared to other methods. Overall, as shown in Fig. 6 (s), (t), (u) and (v), our PRLNet can generate masks with clear boundaries and smooth regions for objects with fine structures, such as *stone fountains* and *leafy plants*. The extensive visualization results in Fig. 6 effectively prove that our method can handle a variety of complex scenarios with superior performance. Above all, the saliency masks generated by PRLNet are consistently the closest to GT.

2) Quantitative Comparison: TABLE II, Fig. 7 and Fig. 8 provide a quantitative comparison of our model with SOTA models on three datasets. First, it can be seen from TABLE II that our PRLNet achieves the highest results on VT821, VT1000 and VT5000. This benefits from the fact that our proposed position-aware relation learning can effectively enhance the intra-class compactness and inter-class separability of feature representations.

Specifically, our PRLNet achieves a marked superiority on VT821. As shown in the results of VT821 in TABLE II, our method improves on average by 0.101, 0.073, 0.152 and 0.043 over the other nine methods for S_α, F_β, E_m and $\mathcal{M}$, respectively. Compared with other methods on VT1000, PRLNet has an average improvement of 0.063, 0.036, 0.094, and 0.026 on the four metrics, respectively. As reported in the results of VT5000 from TABLE II, the performance of our PRLNet has improved by an average of 0.092, 0.067, 0.155, and 0.034 on S_α, F_β, E_m and $\mathcal{M}$, respectively. Moreover, for salient masks, structural similarity (S_α and E_m) can better characterize the homogeneity of foreground-background regions and the sharpness of boundaries. From the above analysis, it can be seen that our PRLNet improves much higher in S_α and E_m metrics than the other two metrics. This indicates that the salient mask of our method is more sophisticated and close to the ground truth. In addition, compared with the previous state-of-the-art method SwinNet [53] on three datasets, our PRLNet achieves an average gain of 1.02%, 1.12%, 0.57%, 13.11% w.r.t S_α, F_β, E_m and $\mathcal{M}$.

Meanwhile, the F-measure curves in Fig. 7 and P-R curves in Fig. 8 also give consistent results. The F-measure combines precision and recall. In Fig. 7, PRLNet usually has achieved remarkable performance under different thresholds.

TABLE III
PERFORMANCE COMPARISON (F-MEASURE, $F_\beta \uparrow$) WITH NINE METHODS ON 13 CHALLENGES OF THE VT5000 DATASET. THE BEST RESULT IS IN BOLD

Methods	BSO	SSO	MSO	CB	CIB	OF	SA	TC	IC	LI	BW	RGB	T
R3Net [38]	0.734	0.538	0.609	0.623	0.654	0.701	0.614	0.608	0.624	0.709	0.562	0.673	0.683
CPD [41]	0.835	0.694	0.765	0.777	0.799	0.801	0.756	0.789	0.764	0.823	0.694	0.804	0.805
PoolNet [42]	0.768	0.624	0.664	0.687	0.717	0.747	0.670	0.686	0.683	0.735	0.661	0.727	0.733
SGDL [51]	0.722	0.715	0.660	0.656	0.654	0.707	0.598	0.621	0.631	0.697	0.583	0.705	0.710
ADF [47]	0.858	0.737	0.806	0.821	0.837	0.806	0.791	0.792	0.803	0.845	0.771	0.840	0.842
FMCF [13]	0.815	0.559	0.724	0.740	0.782	0.743	0.701	0.723	0.725	0.745	0.698	0.762	0.763
MIDD [52]	0.848	0.696	0.781	0.803	0.818	0.799	0.755	0.778	0.768	0.797	0.756	0.817	0.817
ECFFNet [53]	0.878	0.735	0.822	0.840	0.860	0.823	0.801	0.814	0.816	0.850	0.765	0.854	0.855
SwinNet [60]	0.919	0.839	0.882	0.895	0.910	0.890	0.884	0.886	0.875	0.914	0.863	0.903	0.906
Ours	**0.929**	**0.874**	**0.897**	**0.913**	**0.924**	**0.895**	**0.902**	**0.908**	**0.897**	**0.918**	**0.881**	**0.918**	**0.917**

As shown in Fig. 8, our curves noticeably lie above the others on VT821, VT1000 and VT5000 datasets. Our proposed method outperforms the SOTA methods. Above all, both the F-measure curves, P-R curves and quantization results on the three datasets demonstrate the validity and advantages of our PRLNet for RGB-T SOD.

3) Quantitative Comparison on Challenge: To further validate the performance of our PRLNet, we evaluate the performance of each model on all challenges of VT5000 dataset. TABLE I and Fig. 5 summarize the challenges. Challenge-based quantitative comparison results are reported in TABLE III. The best performance of our PRLNet is achieved on all 13 challenges. Compared with SwinNet, our method achieves an average improvement of 1.81% on all challenges. PRLNet achieves an average performance of 0.905 in handling diverse complex targets challenging, such as BSO, SSO, MSO, CB, CIB, OF, SA and TC.

Fig. 6 (b), (c), and (q) show the results on challenges on small objects, multi-object, and large object images, respectively. For example, *soccer goals* and *fences* shown in Fig. 6 (q) and (r) are two common BSO challenges. Compared with the SOTA methods, the BSO mask generated by PRLNet maintains the global consistency of large objects with better intra-class compactness. TABLE III reports that our method achieves the highest performance of 0.929 on BSO. Some of the TC challenges are shown in Fig. 6 (h), (i) and (j), targets such as *ragdolls* on wooden chairs, *water bottles* in bicycle baskets, and *cats* on park seats have similar thermal radiation to their surroundings. The results from Fig. 6 suggest that the existing methods have difficulty in detecting *ragdolls*, *water bottles* and *cats* from the background with TC. In contrast, our PRLNet obviously suppresses the background objects with thermal crossover, and the F-measure attains 0.908 on TC as shown in TABLE III.

The challenges caused by weather or illumination, such as IC, LI and BW, degrade the performance of SOD models. As can be seen from TABLE III, our model still achieves the best performance of about 0.9 for the degraded scenario. In addition, for multispectral RGB-T images, PRLNet effectively learns robust cross-spectral fusion features and reduces the interference caused by spectral inconsistency. The thermal image in Fig. 6 (m) and the RGB image in Fig. 6 (v) have lower quality than the image in the other spectrum. PRLNet overcomes the effect of RGB-T spectral inconsistency and achieves an F-measure above 0.917.

Both the visualization results in Fig. 6 and the quantitative comparisons in TABLE III demonstrate that our method can effectively deal with a variety of salient objects. Above all, the challenge-based quantitative analysis and detailed visualization results consistently demonstrate that our method can effectively address various challenges and outperform state-of-the-art methods.

D. Ablation Study

Our PRLNet mainly contains two key insights: SDM auxiliary module (SDMAM) and feature refinement approach with direction field (FRDF). Therefore, we conduct ablation experiments to verify the validity of components and the involved hyperparameters. Moreover, the intermediate process of the model is visualized to compare the feature maps before and after using different modules. Finally, we summarize the novelties and key differences between the proposed method and existing methods.

1) Effectiveness of SDMAM: SOD methods can be roughly divided into VGG16-based methods [13], [41], [47], [51], [52], ResNet50-based methods [38], [42], [53], and swin transformer (SwinT) based methods [60] according to the different backbones. In TABLE IV, we validate the robustness and effectiveness of SDMAM and FRDF for SOD across different backbone networks, including VGG16, ResNet50, and swin transformer. The corresponding visualization results are shown in Fig. 9 and Fig. 10. The first row of tables represents the baseline model, which does not use the SDMAM and FRDF modules. As can be seen from row 10 in TABLE IV, S_α, F_β, E_m and $\mathcal{M}$ attain 0.916, 0.868, 0.913 and 0.033 on swin transformer, respectively. On the VT5000 datasets, SDMAM improves the performance gain by 7.33% for the SwinT on average across the four metrics. The efficacy of SDMAM module stems from its ability to model the distance relationship between foreground-background pixels and boundaries accurately, thereby enhancing the inter-class separability of salient foreground and background features. Compared with the edge module in SwinNet, TABLE V proves that our proposed SDMAM is more effective and more efficient with fewer parameters and faster speed. As can be observed in Fig. 9 (b) and (d), the separability between foreground

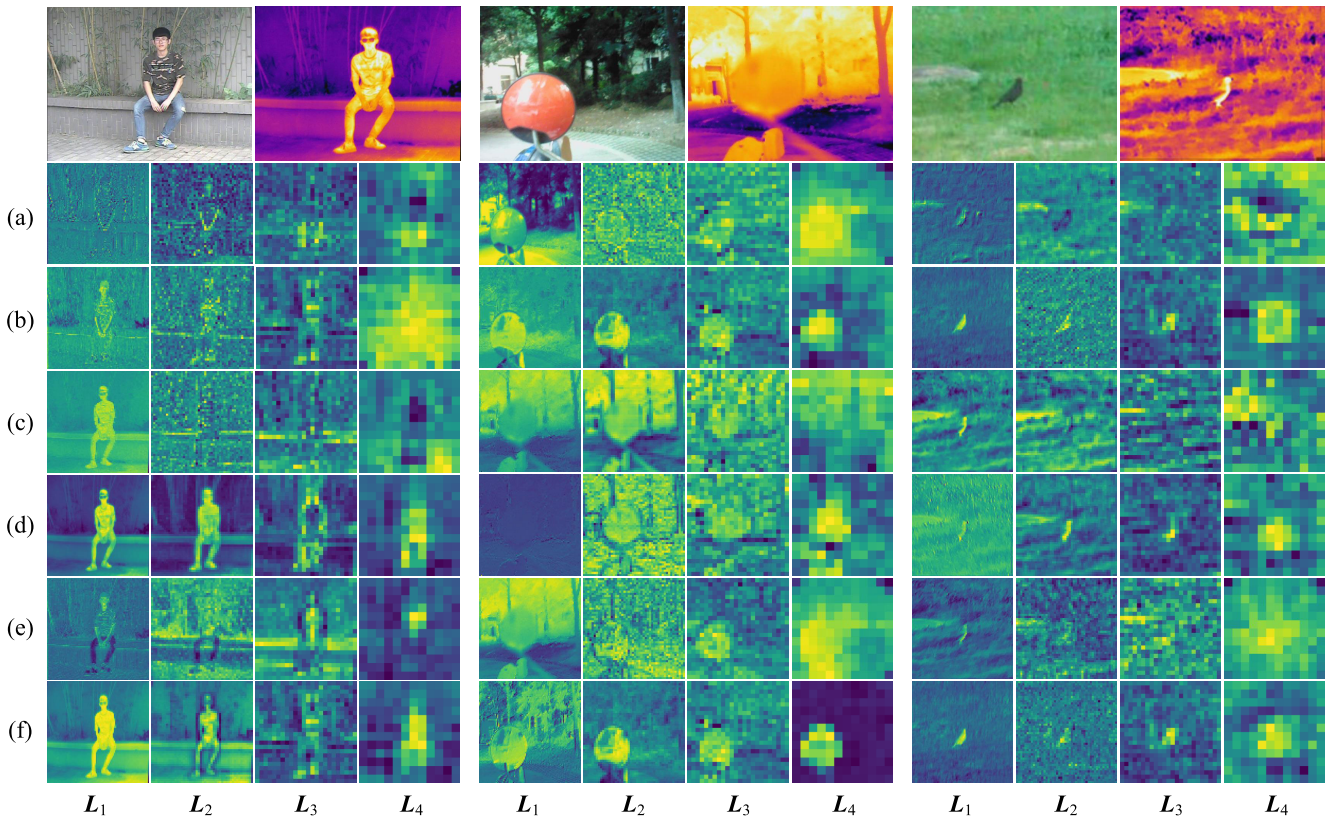


Fig. 9. Visual comparison of the intermediate process features of the proposed module. $L_1 \sim L_4$ denote the corresponding feature maps from low to high level, respectively. (a) and (c) indicate the RGB feature maps and thermal feature maps before SDMAM, respectively. (b) and (d) indicate the RGB feature maps and thermal feature maps after SDMAM, respectively. (e) and (f) indicate the feature maps before and after FRDF, respectively.

TABLE IV
ABLATION ANALYSIS OF SDMAM AND FRDF FOR DIFFERENT BACKBONE NETWORKS.THE BEST RESULTS ARE IN BOLD

Datasets	Modules			VGG16-based [41], [47], [51], *etc.*				ResNet50-based [38], [42], [53]				Swin transformer-based [60]			
	Baseline	SDMA	FRDF	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$
VT821	✓			0.837	0.738	0.859	0.049	0.818	0.713	0.843	0.052	0.838	0.741	0.836	0.047
	✓	✓		0.841	0.749	0.866	0.047	0.819	0.715	0.846	0.051	0.885	0.820	0.888	0.034
	✓		✓	0.845	0.752	0.865	0.045	**0.835**	0.741	0.867	0.048	0.888	0.823	0.893	0.032
	✓	✓	✓	**0.854**	**0.764**	**0.878**	**0.043**	**0.835**	**0.747**	**0.874**	**0.046**	**0.917**	**0.860**	**0.932**	**0.025**
VT1000	✓			0.905	0.858	0.928	0.032	0.820	0.717	0.862	0.037	0.901	0.841	0.899	0.034
	✓	✓		0.914	0.871	0.937	0.028	0.883	0.822	0.905	0.035	0.933	0.895	0.932	0.025
	✓		✓	0.908	0.858	0.930	0.031	0.835	0.741	0.867	0.035	0.939	0.902	0.940	0.022
	✓	✓	✓	**0.911**	**0.866**	**0.932**	**0.030**	**0.896**	**0.840**	**0.918**	**0.034**	**0.944**	**0.902**	**0.951**	**0.016**
VT5000	✓			0.823	0.741	0.859	0.048	0.820	0.717	0.862	0.053	0.904	0.817	0.910	0.042
	✓	✓		0.826	0.747	0.877	0.044	0.818	0.731	0.854	0.044	0.916	0.868	0.930	0.033
	✓		✓	0.826	0.750	0.881	0.043	0.819	0.731	0.852	0.044	0.918	0.866	0.930	0.026
	✓	✓	✓	**0.865**	**0.809**	**0.899**	**0.042**	**0.829**	**0.749**	**0.873**	**0.042**	**0.921**	**0.875**	**0.948**	**0.023**

and background responses is more pronounced after using SDMAM, which points out that SDMAM enhances the inter-class separability of foreground and background.

To further prove the effectiveness and interpretability of our network, we visualize the error maps (*i.e.*, $\boldsymbol{E}_{+SDMAM}$ and $\boldsymbol{E}_{+FRDF}$) of the saliency maps generated by different components. As shown in Fig. 10 (row 6), SDMAM visibly reduces the error pixels and strengthens the separability of inter-class features. The results of $\boldsymbol{E}_{+SDMAM}$ in the Fig. 10 (a) and (d) illustrate that SDMAM notably suppresses the false alarm (*i.e.*, FP). As reported in the ablation experiments, we can conclude that the proposed modules are not just a trick but effective in different approaches.

TABLE V
ABLATION EXPERIMENTS OF DIFFERENT EDGE-AWARE MODULES ON VT5000 DATASETS

Modules	Params(M)	FLOPs(G)	$\mathcal{M} \downarrow$
Edge Module (SwinNet)	0.210	26.96	0.026
SDMAM (Ours)	0.115	18.17	0.023

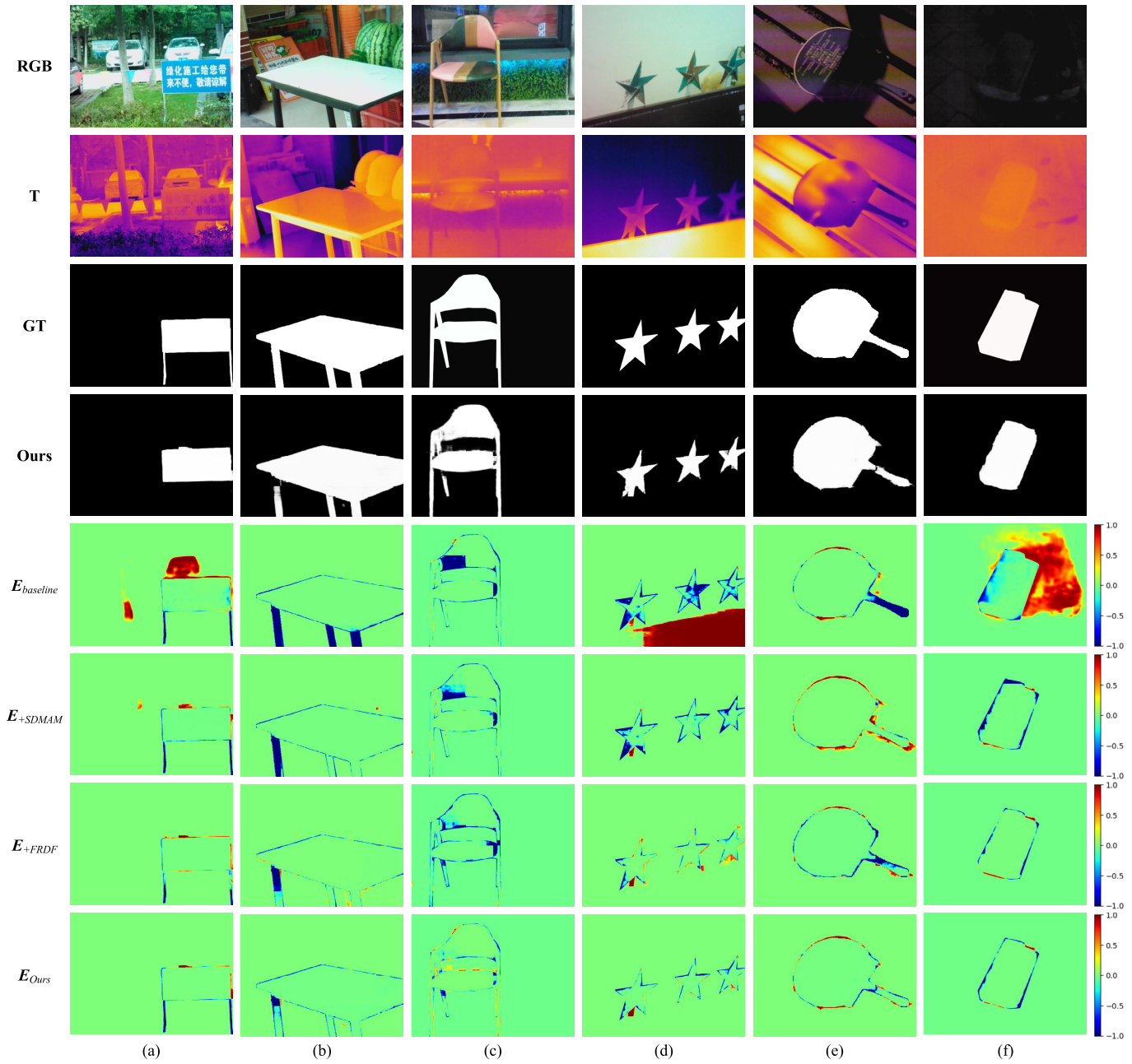
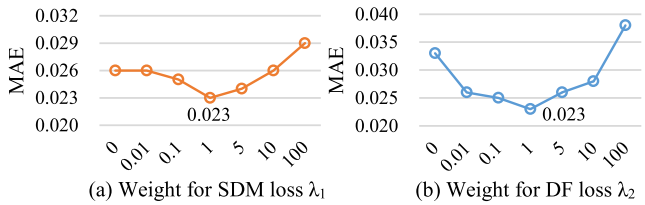


Fig. 10. Visualization results of the error map $\boldsymbol{E} = \boldsymbol{O}_{sal} - \boldsymbol{G}$, $\boldsymbol{E}(p) > 0$ indicate a false positive pixel (FP), *i.e.*, the background is wrongly predicted as an object. $\boldsymbol{E}(p) < 0$ indicates a false negative pixel (FN), *i.e.*, missing some salient target pixels. $\boldsymbol{E}_{baseline}$ represents the error map of the prediction results for the baseline model without SDMAM and FRDF. $\boldsymbol{E}_{+SDMAM}$ and $\boldsymbol{E}_{+FRDF}$ represent the error map after using the SDMAM and FRDF modules, respectively.

2) Effectiveness of FRDF: The FRDF in PRLNet as an auxiliary module rectify the coarse prediction of the decoder. As can be seen from row 11 in TABLE IV, S_α, F_β, E_m and $\mathcal{M}$ for swin transformer attain 0.918, 0.866, 0.930 and 0.026 on VT5000, respectively. FRDF brings an average performance gain of 11.96% for the swin transformer. This suggests that the directional information of object pixels is essential and indispensable for learning a fine feature structure. Furthermore, the comparison of Fig. 9 (e) and (f) shows that the features operated by FRDF have sharper boundaries and are more complete. Based on the above analysis, we verify that FRDF strengthens the intra-class compactness of salient pixels using the direction information between pixels. As shown in Fig. 10 (a), (b), and (e), the error map of the predicted result with FRDF effectively handles the missed detection (*i.e.*, FN) and generates object masks with clear contour and homogeneous regions. The visualization results $\boldsymbol{E}_{+FRDF}$ in Fig. 10 straightforwardly demonstrate the effectiveness of our proposed FRDF. In Section III, we argue that the distance relationship and direction relationship between pixels are crucial for SOD, which can be further proved by this experiment. Above all, we can conclude from TABLE IV, Fig. 9 and Fig. 10 that each component is integral and complementary, which together contribute to the final result. Our proposed



(a) Weight for SDM loss λ_1 (b) Weight for DF loss λ_2

Fig. 11. Hyper-parameters analysis of λ_1 and λ_2 in the position-aware relation learning loss $\mathcal{L}_{prl}$.

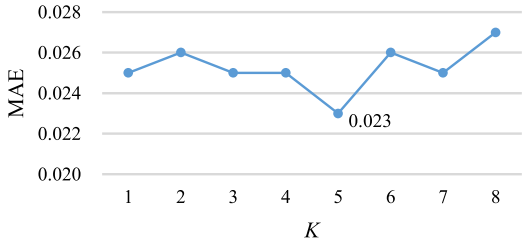


Fig. 12. Hyper-parameters analysis of K in the feature refinement approach with direction field.

SDMAM and FRDF modules can be easily and efficiently integrated with different backbone networks, such as VGG16, ResNet50, and swin transformer, in a plug-and-play manner.

3) Hyper-Parameters Analysis: The parameters λ_1 and λ_2 control the relative importance between SDM loss and DF loss in Eq. (18), which is the decisive hyperparameter for the detection results. The greater the value, the more importance lies in the proposed PRL loss. As shown in Fig. 11 (a), the MAE score decreases as λ_1 grows from 0.01 to 1 and increases as λ_1 grows from 1 to 100, where the valley score reaches 0.023 when $\lambda_1 = 1$. As λ_1 becomes larger, the SDM loss dominates the PRL loss, leading to model performance degradation. The result in Fig. 11 (b) shows that MAE keeps decreasing when λ_2 increases to 1. As λ_2 grows further, the DF loss dominates model training, leading to an increase in the number of error pixels in the saliency mask. Hence, we fix $\lambda_1 = 1$, $\lambda_2 = 1$ in the following experimental settings.

The number of iterations, *i.e.*, K, is another important hyperparameter in our method. As K controls the number of iterations of our proposed FRDF, we conduct experiments to verify the choice of $K = 5$ in Eq. (11). We vary K from 0 to 8, as shown in Fig. 12. The MAE constantly decreases as K is growing from 1 to 5. However, there is a slight increase after 5 due to over-refinement with too many iterations. From the above analysis, we choose $K = 5$ as the number of iterations for our feature refinement approach with a direction field.

V. Conclusion

In this paper, we have proposed a novel position-aware relation learning network (PRLNet) for RGB-T SOD. PRLNet explores the distance and direction relationships between pixels by designing the auxiliary task and optimizing the feature structure to strengthen intra-class compactness and inter-class separation Specifically, we first construct a dual-stream encoder and decoder framework based on swin transformer, where a patch separation layer is designed to upsample the patches in a decoder. Then, we propose SDMAM to learn the distance relationship between foreground-background regions and boundaries, which enhances the boundary perception capability of PRLNet. In addition, we design FRDF to iteratively rectify the features of the bounding pixels using the internal features of the salient objects. FRDF strengthens the intra-class compactness of the salient regions. Extensive experiments and comparisons have shown that the proposed PRLNet consistently outperforms the state-of-the-art methods on three public RGB-T SOD datasets. Notably, the visualization results not only demonstrate that the salient masks generated by PRLNet have sharp boundaries and homogeneous regions, but also offer a new insight to investigate the relationship between pixels. In future work, we will pay more attention to the following two directions: camouflage object detection (COD) and multi-spectral image fusion. Specifically, we will try to apply position-aware relation learning to COD and study the effective complementary fusion between RGB and thermal images. COD needs to effectively perceive the boundaries of objects and generate homogeneous regions. Good multi-spectral image fusion is beneficial for downstream tasks.

References

[1] K. Gu et al., "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.

[2] C. Chen, J. Wei, C. Peng, and H. Qin, "Depth-quality-aware salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 2350–2363, 2021.

[3] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "RGBD salient object detection via disentangled cross-modal fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 8407–8416, 2020.

[4] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1623–1632.

[5] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, and N. Zheng, "Discriminative feature learning with foreground attention for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4671–4684, Dec. 2019.

[6] G. Chen, J. Lu, M. Yang, and J. Zhou, "Learning recurrent 3D attention for video-based person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 6963–6976, 2020.

[7] C. Ma, H. Sun, Y. Rao, J. Zhou, and J. Lu, "Video saliency forecasting transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6850–6862, Oct. 2022.

[8] S. Zhou et al., "Hierarchical and interactive refinement network for edge-preserving salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1–14, 2021.

[9] B. De Brabandere, D. Neven, and L. Van Gool, "Semantic instance segmentation with a discriminative loss function," 2017, *arXiv:1708.02551*.

[10] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, and N. Sang, "Context prior for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12416–12425.

[11] M.-M. Cheng, S.-H. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8006–8021, Nov. 2022.

[12] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3239–3259, Jun. 2022.

[13] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.

[14] Q. Liu et al., "Multi-task driven feature models for thermal infrared tracking," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 11604–11611.

[15] N. Zhang, J. Han, N. Liu, and L. Shao, "Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4167–4176.

[16] Z. Zhu, Z. Zhang, Z. Lin, X. Sun, and M.-M. Cheng, "Co-salient object detection with co-representation purification," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 6, 2023, doi: 10.1109/TPAMI.2023.3234586.

[17] S. Song et al., "Multi-spectral salient object detection by adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12023–12030.

[18] C. Li et al., "Detection-friendly dehazing: Object detection in real-world hazy scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 9, 2023, doi: 10.1109/TPAMI.2023.3234976.

[19] E. Bondi et al., "BIRDSAI: A dataset for detection and tracking in aerial thermal infrared videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1747–1756.

[20] H. Zhou, C. Tian, Z. Zhang, C. Li, Y. Xie, and Z. Li, "PixelGame: Infrared small target segmentation as a Nash equilibrium," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8010–8024, 2022.

[21] Y. Hao, N. Wang, J. Li, and X. Gao, "HSME: Hypersphere manifold embedding for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8385–8392.

[22] H. Zhou, C. Tian, Z. Zhang, Q. Huo, Y. Xie, and Z. Li, "Multispectral fusion transformer network for RGB-thermal urban scene semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[23] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

[24] Z. Zhang et al., "Collaborative boundary-aware context encoding networks for error map prediction," *Pattern Recognit.*, vol. 125, May 2022, Art. no. 108515.

[25] W. Zhou, S. Dong, C. Xu, and Q. Yaguan, "Edge-aware guidance fusion network for rgb-thermal scene parsing," in *Proc. AAAI Conf. Artif. Intell.*, 2022.

[26] X. Zhang, B. Ma, H. Chang, S. Shan, and X. Chen, "Location sensitive network for human instance segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7649–7662, 2021.

[27] W. Zhang, X. Wang, W. You, J. Chen, P. Dai, and P. Zhang, "RESLS: Region and edge synergetic level set framework for image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 57–71, 2019.

[28] Q. Cai et al., "AVLSM: Adaptive variational level set model for image segmentation in the presence of severe intensity inhomogeneity and high noise," *IEEE Trans. Image Process.*, vol. 31, pp. 43–57, 2022.

[29] A. A. Farag, H. E. A. El Munim, J. H. Graham, and A. A. Farag, "A novel approach for lung nodules segmentation in chest CT using level sets," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5202–5213, Dec. 2013.

[30] D. Chai, S. Newsam, and J. Huang, "Aerial image semantic segmentation using DCNN predicted distance maps," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 309–322, Mar. 2020.

[31] L. Lin et al., "BSDA-Net: A boundary shape and distance aware joint learning framework for segmenting and classifying OCTA images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 65–75.

[32] F. Cheng et al., "Learning directional feature maps for cardiac MRI segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 108–117.

[33] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[34] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4884–4893.

[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[36] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020.

[37] L. Zhang, X. Fang, H. Bo, T. Wang, and H. Lu, "Deep multi-level networks with multi-task learning for saliency detection," *Neurocomputing*, vol. 312, pp. 229–238, Oct. 2018.

[38] Z. Deng et al., "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Menlo Park, CA, USA, 2018, pp. 684–690.

[39] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren, "EDN: Salient object detection via extremely-downsampled network," *IEEE Trans. Image Process.*, vol. 31, pp. 3125–3136, 2022.

[40] X. Li et al., "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[41] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3907–3916.

[42] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3917–3926.

[43] C. Li, L. Zhu, G. Tian, Y. Hou, and H. Zhou, "Rethinking referring relationships from a perspective of mask-level relational reasoning," *Pattern Recognit.*, vol. 133, Jan. 2023, Art. no. 109044.

[44] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *Proc. Chin. Conf. Image Graph. Technol.* Singapore: Springer, 2018, pp. 359–369.

[45] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang, "M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Mar. 2019, pp. 141–146.

[46] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2091–2106, Apr. 2022.

[47] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "RGBT salient object detection: A large-scale dataset and benchmark," *IEEE Trans. Multimedia*, early access, May 3, 2022, doi: 10.1109/TMM.2022.3171688.

[48] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "CGFNet: Cross-guided fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2949–2961, May 2022.

[49] F. Huo, X. Zhu, L. Zhang, Q. Liu, and Y. Shu, "Efficient context-guided stacked refinement network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3111–3124, May 2022.

[50] Y. Liang, G. Qin, M. Sun, J. Qin, J. Yan, and Z. Zhang, "Multi-modal interactive attention and dual progressive decoding network for RGB-D/T salient object detection," *Neurocomputing*, vol. 490, pp. 132–145, Jun. 2022.

[51] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 160–173, Jan. 2020.

[52] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive dual-decoder for RGB-thermal salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 5678–5691, 2021.

[53] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1224–1235, Mar. 2022.

[54] N. Zhang, J. Han, and N. Liu, "Learning implicit class knowledge for RGB-D co-salient object detection with transformers," *IEEE Trans. Image Process.*, vol. 31, pp. 4556–4570, 2022.

[55] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[56] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[57] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[58] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[59] C. Zeng and S. Kwong, "Dual swin-transformer based mutual interactive network for RGB-D salient object detection," 2022, *arXiv:2206.03105*.

[60] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4486–4497, Jul. 2022.

[61] H. Zhu, X. Sun, Y. Li, K. Ma, S. K. Zhou, and Y. Zheng, "DFTR: Depth-supervised fusion transformer for salient object detection," 2022, *arXiv:2203.06429*.
[62] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.
[63] Z. Zhang, C. Tian, H. X. Bai, Z. Jiao, and X. Tian, "Discriminative error prediction network for semi-supervised colon gland segmentation," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102458.
[64] Q. Zhang et al., "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, no. 10, pp. 1305–1317, Dec. 2021.
[65] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.
[66] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.
[67] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.
[68] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

Heng Zhou is currently pursuing the Ph.D. degree in electronic science and technology with Xidian University, Xi'an, China. His current research interests include multi-spectral image processing, pattern recognition, and their applications in object detection and segmentation.

Chunna Tian (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in information and communication engineering from Xidian University, Xi'an, China, in 2002, 2005, and 2008, respectively.

From 2006 to 2007, she was a Visiting Student with the Visual Computing and Image Processing Laboratory, Oklahoma State University (OSU). She is currently a Professor with the School of Electronic Engineering, Xidian University. Her current research interests include multimedia analysis, computer vision, pattern recognition, and machine learning. In these areas, she has published around 50 technical articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Pattern Recognition*, *Medical Image Analysis*, and *Neurocomputing*.

Zhenxi Zhang received the B.S. degree in electronic information engineering from Guangxi University, Nanning, China, in 2017, and the Ph.D. degree in electronics science and technology from Xidian University, Xi'an, China, in 2022.

His current research interests include computer vision, medical image processing, and medical image analysis.

Chengyang Li received the M.S. degree in computer technology from the China University of Petroleum (Beijing) in July 2020. He is currently pursuing the joint Ph.D. degree with Peking University (PKU) and AMS.

He has published many papers in SCI journals and top conferences. His research interests include image processing, video understanding, and multimodal intelligence.

Yuxuan Ding received the B.Eng. degree in intelligent science and technology from Xidian University, Xi'an, China, in 2018, where he is currently pursuing the Ph.D. degree in information and communication engineering.

His main research interests include machine learning, computer vision, vision-language, and their applications.

Yongqiang Xie is currently a Researcher with the Institute of Systems Engineering, AMS, and a Professor with the School of Systems Science and Engineering, Sun Yat-Sen University. He is the Leader of the dedicated algorithm group for China's video and audio standards. He has been engaged in the research of image, video, and communication technology for a long time, and has made outstanding contributions to signal processing. He is undertaking more than 20 major national scientific research projects. He has published two books and around 50 technical papers in refereed journals and proceedings. He holds more than ten granted patents. His current research interests include visual information processing, multimedia analysis, machine learning, and pattern recognition. He is a member of the China Postdoctoral Fund Review Committee. He received the Qiu Shi Award and was selected as a Candidate for the One Hundred Plus One Thousand Plus Ten Thousand Talents Project of the New Century and Outstanding Mid-Aged Expert. He was selected as an Expert enjoying the Government's Special Subsidy. He is the Executive Deputy Director of the Rich Media Committee of the Chinese Institute of Command and Control.

Zhongbo Li is currently a Senior Engineer with the Institute of Systems Engineering, AMS. He is mainly dedicated to video understanding and intelligent analysis, including pedestrian detection, crowd counting, and intelligent transportation. He has published more than 20 high-level articles in related fields and published one academic book.

He received the Second Prize in the National Science and Technology Progress Award and the First Prize in the Provincial and Ministerial Science and Technology Award.