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.

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].
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 focuses 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. Then, we construct SDMAM for encoders to learn the distance relationship between regional pixels and boundary pixels. Next, the reverse swin transformer decoder aggregates the complementarity between different levels of RGB-T features, 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. Finally, we propose a novel position-aware relation learning loss to generate object masks with clear boundaries and homogeneous regions.

C. Swin Transformer

Compared with CNN, transformer has an advantage in modeling long-range dependencies and achieving 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 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. Liu et al. 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. 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 strengthen the salient object pixel set and background pixel set, respectively. SDM not only perceives the boundary of an object but also looks inside (i.e., $D(p, b) = \text{SDM}(G, p) < d$) or outside (i.e., $D(p, b) = \text{SDM}(G, p) > d$) the boundary. As shown in Fig. 1 (a), SDM transformation $D(p)$ for each pixel $p \in G$ is given by:

$$D(p) = \begin{cases} - \inf_{b \in \partial S} d(p, b), & p \in S_{\text{sal}} \\ 0, & p \in \partial S \\ + \inf_{b \in \partial S} d(p, b), & p \in S_{\text{bg}} \end{cases} \quad (3)$$

where $\inf$ denotes the infimum, $b$ is the boundary pixel. In Eq. (3), $\partial S$ is the zero level set which represents the pixel set of the target boundary. $S_{\text{sal}}$ and $S_{\text{bg}}$ 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 $p \in G$, the sign of $D(p)$ indicates whether it is located outside (i.e., $D(p) > 0$) or inside (i.e., $D(p) < 0$) the object. $D(p) = 0$ denotes the boundary of the object. $|D(p)|$ 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

**Algorithm 1.**



**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.

**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.
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,

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

where $x^i = \text{concat}(x^i_1, x^i_2)$, $i = 1, 2, 3$. $D \in \mathbb{R}^{h \times w \times 1}$ represents the prediction result of SDMAM. The dimensions of $x^1, x^2$ and $x^3$ 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 $\{x^1, x^2, x^3\}$ are further fused by pointwise convolution (PWConv) [62] with ReLU and upsampling operations,

$$y^i = \left[\text{PWConv}\left(x^j\right)\right]^{\times(2^{i-1})},$$  \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 $y^i \in \mathbb{R}^{\frac{h}{2^i} \times \frac{w}{2^i} \times 32}$. Finally, the multi-scale multi-spectral features $y$ are fused by $3 \times 3$ convolution and upsampled to the resolution of the input image.

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

where the output of SDMAM is $D \in \mathbb{R}^{h \times w \times 1}$, Conv$_{3 \times 3}$ indicate $3 \times 3$ 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 $x^i = \text{concat}(x^i_1, x^i_2)$ is flattened into an input sequence. A set of queries $Q$, keys $K$, and values $V$ is computed by embedding the input sequence into three weight matrices.

In Eq. (7), we compute cross-spectral attention $z^i$ as in [22] and [55].

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

where $z^i \in \mathbb{R}^{\frac{h}{2^i} \times \frac{w}{2^i} \times 8c}$, $\sqrt{d}$ 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).

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

where $i = 1, 2, 3$. The dimensions of $z^1, z^2$ and $z^3$ are $\mathbb{R}^{\frac{h}{8} \times \frac{w}{8} \times c}$, $\mathbb{R}^{\frac{h}{16} \times \frac{w}{16} \times 2c}$ and $\mathbb{R}^{\frac{h}{32} \times \frac{w}{32} \times 4c}$, respectively. The salient swin transformer decoder output $z \in \mathbb{R}^{h \times w \times 64}$ is obtained by upsampling $z^1$ 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 $F$ is shown in Eq. (9). The direction of $F(p)$ is from $b$ pointing to $p$, and $b$ is the nearest pixel to $p$ on the boundary. For the pixel $p \in G$,

$$F(p) = \begin{cases} 
    \inf_{\forall \theta \in \partial S} b\theta, & p \in S_{\text{sal}} \\
    (0, 0), & p \in S_{\text{bg}}
\end{cases}$$  \tag{9}

where $S_{\text{sal}}$ and $S_{\text{bg}}$ 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 $z$ to predict the direction field feature $F \in \mathbb{R}^{h \times w \times 2}$ in Eq. (10).

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

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

$$z_k(p) = z_k\left(p_x + F_x(p), p_y + F_y(p)\right),$$  \tag{11}

where $z_k$ 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 $z^{*} \in \mathbb{R}^{h \times w \times 2c}$. 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 $z$ with the rectified feature $z^*$ to generate the salient mask.
$O_{\text{sal}} \in \mathbb{R}^{h \times w \times 1}$. Both SDM and DF are learned in a fully-supervised way, which will be discussed in Sec. III-E. Based on the ground truth $G$, we obtain the true supervised signal for the SDMAM and FRDF.

E. Loss Function

According to the ground truth $G$ 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).

$$D_{gt} = \mathcal{D}(G),$$
$$F_{gt} = \mathcal{F}(G),$$

(12)

In Eq. (12), $D_{gt}$ is the true SDM and $F_{gt}$ 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

$$L_{sdm} = \sum_{p \in \Omega} \|D - D_{gt}\|^2,$$

(13)

where $\Omega$ denotes all pixels, $D$ is the predicted result of SDMAM. $L_{sdm}$ 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

$$L_{df} = \sum_{p \in \Omega} \left( \|F - F_{gt}\|_2 + \|\cos^{-1}(F, F_{gt})\|^2 \right),$$

(14)

where $F$ and $F_{gt}$ indicate the predicted DF and the corresponding ground truth, respectively. $L_{df}$ 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 ($L_{DS}$) 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]. $L_{DS}$ is defined as follows,

$$L_{DS}(O, G) = \sum_{p \in \Omega} \sum_{x, y} w(p) \psi \left( |\partial_x O| e^{-\alpha |\partial_y O|} \right),$$

(15)

where

$$w(p) = \left\{ \begin{array}{ll} \|\mathcal{F}(p)\|^{-1}, & p \in S_{sal} \\ 1, & p \in S_{bg} \end{array} \right.$$ \hspace{1cm} (16)

$$\psi(m) = \sqrt{m^2 + 0.01}^2.$$  

$O$ and $G$ 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), $w(p)$ indicates the weight on pixel $p$. Therefore, the saliency loss is

$$L_{sal} = L_{DS}(O_{\text{sal}}, G).$$

(17)

Finally, our position-aware relation learning loss $L_{prl}$ is

$$L_{prl} = L_{sal} + \lambda_1 L_{sdm} + \lambda_2 L_{df},$$

(18)

where $\lambda_1$ and $\lambda_2$ 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 VTR821 [44], VT1000 [51] and VT5000 [47], which have 821, 1000, and 5000 aligned image pairs, respectively. Compared with VTR821, 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

Algorithm 1 The Pipeline of PRLNet

1. **Input:** RGB-T images $\{I_r, I_t\}$, ground truth $G$

2. **Output:** Salient object mask $O_{\text{sal}}$, signed distance map $D$, direction field $F$

3. 

4. **while** epoch $< N$ **do**

5. 

6. 

7. 

8. 

9. **for** $k \leq K$ **do**

10. 

11. 

12. 

13. 

14. 

15. 

16. 

17. 

18. **end while**

19. **return** $O_{\text{sal}}$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
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. |

The whole 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.

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.

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).

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 $(M \downarrow)$ [68]. $\uparrow$ and $\downarrow$ indicate that the higher the better and the lower the better, respectively. The P-R curves and $F\beta$ evaluate the quality of the prediction results in terms of Precision and Recall. $S\alpha$ and $E_m$ mainly measure the structural similarity between the predicted saliency mask and GT. $M$ 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).

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}.$$  \hfill (19)

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

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

$$S\alpha = \alpha S_o + (1 - \alpha) S_r,$$  \hfill (20)

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

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

$$F\beta = \frac{(1 + \beta^2) \times P \times R}{\beta^2 \times P + R},$$  \hfill (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].

$$E_m = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \phi_{ij}.$$  \hfill (22)

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

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

$$M = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |O_{ij} - G_{ij}|.$$  \hfill (23)
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 (g)-(i)), 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) and (n)) 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
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. From 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 $S_{\alpha}$, $F_{\beta}$, $E_{m}$ and $M$, 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 $S_{\alpha}$, $F_{\beta}$, $E_{m}$ and $M$, respectively. Moreover, for salient masks, structural similarity ($S_{\alpha}$ 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 $S_{\alpha}$ 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 $S_{\alpha}$, $F_{\beta}$, $E_{m}$ and $M$.

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.
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 0.9 in handling diverse complex targets challenging, such as BSO, SSO, MSO, CB, CIB, OF, SA and TC.

![Table III](image)

| 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 |
| SSDL [51] | 0.722 | 0.715 | 0.660 | 0.656 | 0.654 | 0.704 | 0.598 | 0.621 | 0.631 | 0.697 | 0.583 | 0.705 | 0.710 |
| ADF [47] | 0.838 | 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 |
| MDD [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 | 0.894 | 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 |

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, $S_\beta$, $F_\beta$, $E_m$ 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.

| Datasets | Modules | VGG16-based [41], [47], [51], etc. | ResNet50-based [38], [42], [53] | Swin transformer-based [60] |
|----------|---------|-----------------------------------|----------------------------------|---------------------------|
|          | Baseline | $S_a \uparrow$ | $F_p \uparrow$ | $E_m \uparrow$ | $M \downarrow$ | $S_a \uparrow$ | $F_p \uparrow$ | $E_m \uparrow$ | $M \downarrow$ | $S_a \uparrow$ | $F_p \uparrow$ | $E_m \uparrow$ | $M \downarrow$ |
| VT821    | ✓        | 0.837 | 0.738 | 0.859 | 0.049 | 0.818 | 0.713 | 0.843 | 0.052 | 0.858 | 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 | 0.883 | 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 |
|          | ✓        | ✓      | ✓      | ✓      | 0.911 | 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., $E_{+SDMAM}$ and $E_{+FRDF}$) 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 $E_{+SDMAM}$ 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**

| Modules | Params(M) | FLOPs(G) | $M \downarrow$ |
|---------|-----------|----------|---------------|
| 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 $E = O_{sal} - G$. $E(p) > 0$ indicate a false positive pixel (FP), i.e., the background is wrongly predicted as an object. $E(p) < 0$ indicates a false negative pixel (FN), i.e., missing some salient target pixels. $E_{baseline}$ represents the error map of the prediction results for the baseline model without SDMAM and FRDF. $E_{+SDMAM}$ and $E_{+FRDF}$ 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, $S_\alpha$, $F_\beta$, $E_m$ and $M$ 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 $E_{+FRDF}$ 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
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.

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., “Salient-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, “RGBDB salient object detection via disentangled cross-modal fusion,” IEEE Trans. Image Process., vol. 29, pp. 8406–8416, 2020.
[4] M. Feng, H. Lu, and E. Ding, “Attribute 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] J. Wang, 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.
[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. Pattern Recognit., Jun. 2009, pp. 1597–1604.

[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.

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 multimedia analysis, computer vision, 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.

He is currently a Professor with the School of Electronic Engineering, Xidian University. His 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.

Chengyang Li received the B.S. degree in computer science 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.

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.