DPPNet: A Depth Pixel-wise Potential-aware Network for RGB-D Salient Object Detection

Junbin Yuan, Yiqi Wang, Zhoutao Wang, Qingzhen Xu, Bharadwaj Veeravalli, *Senior Member, IEEE*, and Xulei Yang, *Senior Member, IEEE*

***Abstract*—Depth cues are essential for visual perception tasks like Salient Object Detection (SOD). Due to varying depth reliability across scenes, some researchers propose evaluating the overall quality of the depth maps and discarding the less reliable ones to avoid contamination. However, these methods often fail to fully utilize valuable information in depth maps, leading to sub-optimal performance particularly when depth quality is unreliable. Since low-quality depth maps still contain useful information that potentially improves model performance, we propose a Depth Pixel-wise Potential-aware Network to leverage these depth cues effectively. This network includes two novel components designed: 1) A learning strategy for explicitly modeling the confidence of each depth pixel to assist the model in locating valid information in the depth map. 2) A cross-modal adaptive multiple fusion module that fuses features from both RGB and depth modalities. It aims to mitigate the contamination effect of unreliable depth maps and fully exploit the benefits of multiple fusion strategies. Experimental results show that on four publicly available datasets, our method outperforms 17 mainstream methods on various evaluation metrics. The code is available at: https://github.com/danielfaster/DPPNet.**

***Index Terms*—RGB-D salient object detection, Pixel-wise depth potential-aware, Multi-modality fusion.**

I. Introduction

The goal of salient object detection (SOD) is to locate interesting regions that attract human attention most in an image or video. As a pre-processing technique, SOD benefits a lot of vision tasks, such as object segmentation [1], medical segmentation [2], object tracking [3] and video analysis [4], [5]. In the past years, due to their powerful feature expression capabilities, CNN-based SOD methods have made significant advances in performance and have become mainstream techniques.

However, RGB methods often struggle in challenging scenarios, such as those with complex backgrounds or low contrast between the foreground and background. In contrast, depth maps provide rich geometric information that can be used to distinguish foreground and background targets from [7], [8] proposed complex backgrounds. Inspired by this, some researchers incorporating depth information into the SOD process. They found that by designing a fusion scheme that exploits the complementarity of the visual information of RGB and the spatial information of the depth map, the model can better locate salient regions in challenging scenes. Despite the proliferation of multimodal fusion schemes that show promising performance, many of them do not consider the reliability of the depth maps that vary with the scene. If unreliable depth maps are directly fused, contamination is likely to affect detection results [9]–[11]. To address this issue, some researchers have proposed some measures of depth map reliability and advocated either directly discarding low-quality depth maps or adjusting the fusion scheme according to the quality of the depth map [9], [10]. Fan et al. [9] proposed to evaluate the quality of the depth map by calculating the distance between the prediction maps generated by the RGB and RGB-D branches, and directly discarding the depth map when the result is unreliable. Chen et al. [10] adjusted the fusion scheme according to the quality score of the depth map, which is obtained by calculating the ratio of the intersection

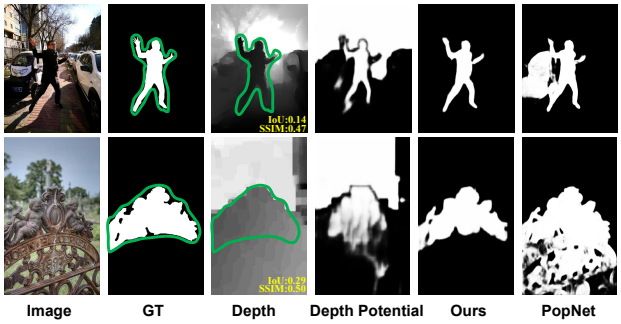

Fig. 1: Despite the lower overall quality scores of unreliable depth maps (measured by Structural Similarity Index (SSIM) and Intersection over Union (IoU) between the ground truth and binary depth maps), they still contain valid depth regions that are useful for detection (as indicated by the green bounding boxes (BB)). With the help of the depth potential map, our approach excels in extracting valid depth information from unreliable depth maps, outperforming the state-of-the-art (SOTA) method [6]. The depth potential value represents the confidence of each point in the depth map generated by our model, with whiter colors indicating higher confidence levels at those points. The green BB is the valid depth region within the depth map.

The work was supported in part by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021) and in part by the International Training Program for Young Excellent Researchers in Guangdong Province. The corresponding author is Qingzhen Xu.

Junbin Yuan, Yiqi Wang, and Qingzhen Xu are with School of Computer Science, South China Normal University, Guangzhou, 510631, China
Zhoutao Wang, College of Computer Science and Technology of Nanjing University of Aeronautics and Astronautics, NanJing, 211106, China
Bharadwaj Veeravalli, Electrical and Computer Engineering, National University of Singapore (NUS), 117574, Singapore
Xulei Yang, Institute for Infocomm Research (I^2R), A*STAR, 138542, Singapore

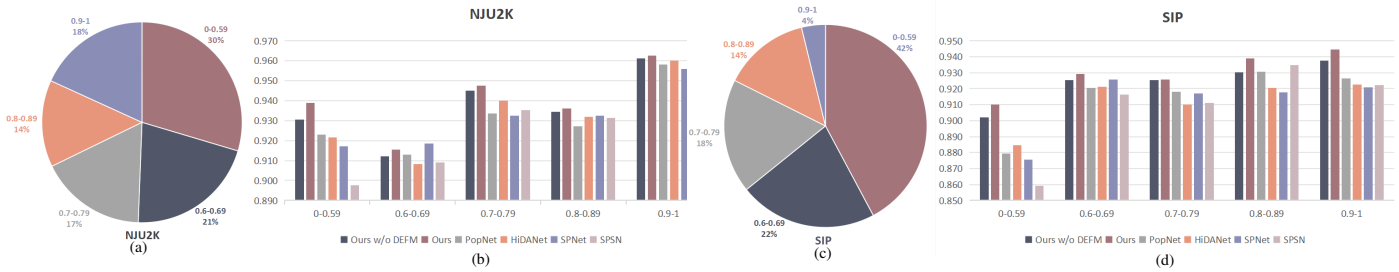

Fig. 2: (a) and (c) represent the distribution of various similarity scores between ground truth and binary depth maps across the NJU2K [12] and SIP [9] datasets. A higher similarity score (calculated through SSIM [13]), indicates greater similarity between images, suggesting higher reliability of the depth map [10]. (b) and (d) depict the performance of some mainstream methods (PopNet [6], HiDANet [14], SPSN [15], and SPNet [16]) on different similarity images. The horizontal axis represents five different similarity intervals, while the vertical axis indicates the scores of the different methods on the max F-measure [17] metric.

area between the binary depth map and the ground truth. These methods have produced promising results. However, the overall assessment for depth map quality is difficult to identify valid information within unreliable depth maps containing significant noise. Therefore, these methods often fail to fully exploit the valuable information in depth maps, especially when the depth quality is unreliable (see Fig.1). This is likely to limit the potential performance of the RGB-D model, as low-quality depth maps still contain useful information that can improve model performance.

As demonstrated in Fig. 2 (a) and (c), the number of unreliable depth maps is significant in two prominent datasets. Additionally, most mainstream models exhibit sub-optimal performance when presented with these unreliable depth maps, as shown in Fig. 2 (b) and (d). Therefore, how to extract effective depth information from unreliable depth maps for more accurate detection is valuable research. To tackle the aforementioned challenges, a depth pixel-wise potential-aware network is proposed that achieves the localization of effective information regions in the depth map. To explicitly model the confidence of each depth pixel, a learning strategy is proposed to assist the model in locating valid information in the depth map. The comparison between "Ours" and "Ours w/o DFEM" is shown in Fig. 2 (b) and (d), it indicates that the depth potential map can be utilized to refine the depth map and facilitate performance improvement. Moreover, we incorporate a cross-modal adaptive multiply fusion module to enhance the focus on pertinent depth information, thereby maximizing the utilization of cross-modal complementarity and preventing contamination. As shown in Fig. 1, with the assistance of the depth potential map, our model accurately locates the valid regions within the depth map, even performing better in cases where the depth map is unreliable. In real-world scenarios, external factors often degrade the quality of depth maps, making them less reliable. Our method offers a novel approach to effectively utilize these low-quality depth maps, enhancing detection accuracy. This capability is particularly crucial for applications such as autonomous driving, robotics, and augmented reality.

Our approach contains the following main contributions:

- We propose a Depth Pixel-wise Potential-aware Module (DPPM). Unlike the conventional approach of using a single score to assess the entire depth map's reliability, DPPM evaluates the potential of each individual pixel in the depth map, which helps to fully exploit the effective depth information to better identify the salient objects, especially when the depth map is unreliable.
- We develop a cross-modal adaptive multiple fusion module (CAMFM). To guide the model to better focus on effective depth information, we design an adaptive fusion scheme for dynamic depth perception. In addition, we use multiple fusion strategies to fully exploit the complementary nature of different modality data.
- We present a depth pixel-wise potential-aware network (DPPNet) that perceives valid depth regions in the depth map and effectively integrates the complementarity from the cross-modal RGB-D data, leading to better performance. Experimental results show that our model achieves SOTA performance on multiple publicly available datasets compared to 17 mainstream models.

II. Related Work

In this section, we will review the salient object detection models for RGB and RGB-D images, with a particular focus on deep learning-based approaches. These methods have made impressive progress in recent years.

A. RGB Salient Object Detection

Salient object detection primarily aims to local and segment regions of visual saliency in images. In the past two decades, hundreds of hand-craft feature methods have been proposed. Most of these works are based on manual features (such as center prior and contrast prior) called traditional methods. More details of them can be seen in [18].

Convolutional neural networks (CNNs) have become popular due to their powerful feature extraction capabilities that have set new best records for many computer vision tasks. While these methods have made great progress in terms of performance compared to traditional methods, they still suffer from high computational costs. Lately, inspired by the successful application of Fully Convolution Network (FCN) in

semantic segmentation [19], FCN-based methods have gradually become mainstream. Wang et al. [20] proposed a recurrent FCN for SOD method. To deal with the scale-space problem, Hou et al. [21] proposed short connections to the skip-layer structures within the HED [22] architecture. To fully leverage inter-image information, Jerripothula et al. [23] proposed a quality metric for assessing saliency map quality, which guides saliency fusion to achieve more accurate detection. In order to obtain clearer saliency contours, Zhao et al. [24] proposed a multi-task model for explicit edge modeling by adding an edge detection task to obtain more edge features to complement the spatial features of the salient objects. Inspired by the logical interrelationship between the binary segmentation graph and the edge graph, Wu et al. [25] propose a network called SCRN that optimizes salient features and edge features in both directions. Inspired by the strong consistency between fixation prediction and salient object detection, Yuan et al. [26] used the fixation as a kind of attention information to guide both saliency detection and boundary detection, resulting in more accurate localization and complete segmentation of the salient object. To exploit different types of saliency cues, Jerripothula et al. [27] explored a multi-saliency cue fusion method based on the appearance domain and proposed a novel appearance fusion framework to effectively combine these cues. For exploring the potential of pooling techniques in convolutional neural networks, Liu et al. [28] proposed PoolNet+ for SOD and obtained a more accurate location of salient objects with sharpened details. To fully utilize the advantages of ViT in computer vision demonstrations, several transformer-based models have been proposed. Liu et al. [29] presented a prominence detection method based solely on transformers. Yuan et al. [30] proposed a dual-model parallel architecture that combines CNN and transformers to exploit the complementary advantages of CNN's local modeling and transformers' global context capturing. While the above methods have made great progress, there is still room for improvement in challenging scenarios.

B. RGB-D Salient Object Detection

Over the past several years, many RGB-D SOD models have been developed. Similarly to RGB SOD, the development of RGB-D models has evolved from traditional methods [7], [12] to deep learning-based methods. This section will mainly discuss the deep models and more details of the traditional method can be found in [31]. Over the past few years, numerous studies have employed deep neural networks (DNNs) to fuse RGB-D data in order to design deep models and have made significant advancements. Qu et al. [8] first introduced CNNs into the RGB-D SOD task. Chen et al. [32] designed a novel complementarity-aware fusion module to better capture the complementary part between cross-modal and cross-level features. Cong et al. [33] propose PICR-Net, which achieves a cross-attention mechanism for multi-modality fusion at a lower cost, effectively leveraging the complementary nature of different modalities.

To effectively explore the correlations between RGB images and depth maps, several methods [34], [35] have focused on exploiting effective multi-modal correlations and multi-scale information to boost SOD performance. Chen et al. [36] proposed a multi-scale multi-path fusion network with cross-modal interactions (MMCI) that diversify the fusion pathways and introduce cross-modal interactions in multiple layers to explore complementarity between multi-modal and multi-level feature representations. Fu et al. [37] proposed a joint learning and densely cooperative fusion architecture. It consists of a joint learning module that conducts joint learning for the two modalities to discover the commonality between them and a densely cooperative fusion module that exploits the complementarity between RGB and depth modalities from a feature-based viewpoint. Cong et al. [33] proposed a CNN-based Transformer architecture to comprehensively model global long-range dependencies both within and between modalities. They also introduced a method that uses positional constraints to integrate complementary information from features of different modalities. Zhu et al. [38] rethought the differences and correlations between encoding and decoding features and proposed the Cross-Modal Inverse Guidance Network (CMIGN) to fully exploit these features for SOD. Wu et al. [39] proposed an asymmetric network for RGB-D SOD to capture the long-range multi-modality interactions for feature fusion and explore intra- and inter-pixel relationships through a pixel-level contrastive learning method. Lin et al. [40] proposed a multi-branch architecture that uses separate branches for encoding and decoding each modality to better preserve their details. An additional branch is used for multi-scale fusion of multi-modal features and saliency detection.

These methods demonstrated the advantages of integrating RGB and depth information; however, they failed to consider the detrimental effects of inaccurate or blurred depth maps on detection performance. Consequently, various approaches [9]–[11], [41], [42] have been proposed to address this issue and mitigate its adverse effects. Cong et al. [11] argued the quality of depth maps is important and proposed a depth confidence measure to evaluate the reliability of depth maps to alleviate the negative impacts caused by low-quality depth maps. Zhao et al. [41] advocated the importance of high-quality depth maps and introduced a contrast loss for depth map enhancement. Chen et al. [10] introduced DPANet, incorporating a depth potential perception module. This module predicts the confidence of the depth map and guides cross-modal fusion, mitigating contamination effects from unreliable depth information. Fan et al. [9] proposed a depth depurator unit to explicitly judge whether to filter low-quality depth maps by comparing the resulting output by the RGB-D and depth subnetworks. Ji et al. [43] introduced a depth calibration strategy to identify poor-quality depth maps and generate a calibrated depth map, thus addressing the depth quality challenge. Wang et al. [44] proposed a multi-level depth quality measurement that devised three features to measure depth quality to guide selective cross-modal fusion. Chen et al. [45] introduced a novel DCA (Depth Contribution Assessment) subnet to predict the "Most Valuable Depth Regions," which could potentially enhance the performance of the RGB saliency subnet. Zhang et al. [46] proposed an efficient lightweight RGB-D SOD

model called DFM-Net. The model employs an efficient Depth Quality-Inspired Feature Manipulation (DQFM) process that explicitly avoids the introduction of noisy depth features and enhances depth features during cross-modal fusion, resulting in improved detection accuracy. Xiao et al. [47] proposed an algorithm to enhance depth maps and reduce the impact of low-quality depth maps. The methods mentioned above attempt to address the negative impact of unreliable depth maps on models and propose various solutions. However, the challenge of effectively extracting useful depth information from such unreliable maps remains under-explored.

III. Methodology

A. Overview

As shown in Fig. 3, the proposed network is a two-stream architecture, and a multi-branch structure is employed to execute multi-task prediction. To explore the shared information between RGB and depth images while preserving the features specific to each modality, two parallel branches were employed to generate independent saliency maps for RGB and depth data. Subsequently, another branch was utilized for cross-modal fusion to produce the saliency map, serving as the saliency prediction outcome of the model. Furthermore, two additional branches were designed for edge prediction and depth potential prediction, respectively. The specific process is as follows:

First, the RGB images I_{rgb} and depth images I_{depth} are fed separately to the Transformer and CNN. Second, a Feature Alignment Module (FAM) [30] is employed to transform the output of the Transformer to obtain multi-level feature representations to align the output of CNN. Thirdly, four feature converters, each comprising three convolutional layers, are employed to extract features for the four tasks: saliency prediction, edge prediction, depth prediction, and depth potential prediction. Fourthly, a Depth Pixel-wise Potential-aware Module (DPPM) is proposed to perceive the reliability of each point in the depth map. Furthermore, a Cross-modal Adaptive Multiple Fusion Module (CAMFM) is designed to perform feature enhancement and adaptive feature fusion. Finally, five decoders with typical U-Net structures and the receptive field block [48] are incorporated to obtain prediction maps. The proposed network saliency prediction results are derived from the decoding result of F_s, which represents the fusion result of the RGB and depth modalities, effectively leveraging the complex correlations between the two modalities.

B. Depth Pixel-wise Potential Perception

To alleviate the negative impact of the unreliable depth map, Fan et al. [9] proposed to measure the reliability of the depth map and discard the unreliable one. Chen et al. [10] proposed to perceive the potential of depth maps and adjust the fusion process based on their reliability to mitigate the negative impact of unreliable depth. However, we find that there are still some valid depth regions in the unreliable depth map that can promote salient object detection (see Fig. 1).

Based on this finding, we argue that extracting valid features from the depth map is significant in boosting the model performance. Consequently, we design a deep pixel-wise potential-aware network for measuring the confidence level of each point in the depth map and guiding the model to focus on the useful features. Different from the work [10], our model assesses the potential of the depth map with a finer granularity. Specifically, DAPNet [10] evaluates the whole depth map whereas our proposed model measures the confidence level of each point in the depth map, resulting in a more accurate valid region for capturing.

For the purpose of attaining more precise predictive outcomes, we advocate training the depth potential perception network through a supervised approach. However, a challenge we faced was the absence of any depth potential label in the most mainstream datasets. The reliability of the depth map can be assessed by measuring the similarity between the binary depth map and its corresponding ground truth [10]. Inspired by this, we generated the depth potential perception labels. The generated label is an image in which each point represents the similarity of the corresponding area between the binary depth map and the ground truth. The greater the similarity, the higher the pixel value in the depth potential label, which indicates the higher reliability of the depth in that region. Therefore, the color in the depth potential label represents its reliability. The whiter the colour, the more reliable it is, and vice versa.

The processes for the generation of the depth potential labels and the depth potential maps are as follows:

1) Generation of Depth Potential Labels: Firstly, the depth map is binarized. Since the truth map I_{gt} for significance testing is a binary map and the depth map I_{depth} is a greyscale map, we binarize the depth map using Otsu's [49] method.

Secondly, we use a window with a preset size of $K \times K$ to extract the content descriptions of the two identical regions separately (the formula for the procedure is given in (1)).

$$\begin{aligned} R_{i,j} &= I_{gt}[i : i+K, j : j+K], \\ Q_{i,j} &= I_{depth}[i : i+K, j : j+K], \end{aligned} \tag{1}$$

where $R_{i,j}$ and $Q_{i,j}$ represent the content extracted from the images I_{gt} and I_{depth} at position (i, j) with a window size of $K \times K$, respectively. We set K to be 8.

Thirdly, the SSIM [13] is used to calculate the similarity between these two regions, described as follows (2).

$$\text{SSIM_Map}_{i,j} = \text{SSIM}(R_{i,j}, Q_{i,j}), \tag{2}$$

where $\text{SSIM_Map}_{i,j}$ is the SSIM value between the region $R_{i,j}$ in I_{gt} and the corresponding region $Q_{i,j}$ in I_{depth}.

SSIM is defined as follows (3):

$$\text{SSIM}(R_{i,j}, Q_{i,j}) = \frac{(2\mu_{R_{i,j}}\mu_{Q_{i,j}} + c_1)(2\sigma_{R_{i,j}Q_{i,j}} + c_2)}{(\mu_{R_{i,j}}^2 + \mu_{Q_{i,j}}^2 + c_1)(\sigma_{R_{i,j}}^2 + \sigma_{Q_{i,j}}^2 + c_2)}, \tag{3}$$

where $\mu_{R_{i,j}}$ and $\mu_{Q_{i,j}}$ are the mean luminance, $\sigma_{R_{i,j}}$ and $\sigma_{Q_{i,j}}$ are the standard deviations, and $\sigma_{R_{i,j}Q_{i,j}}$ is the covariance between images I_{gt} and I_{depth}. Constants c_1 and c_2 are included for numerical stability. The index ranges from -1 to 1, where 1 indicates identical images.

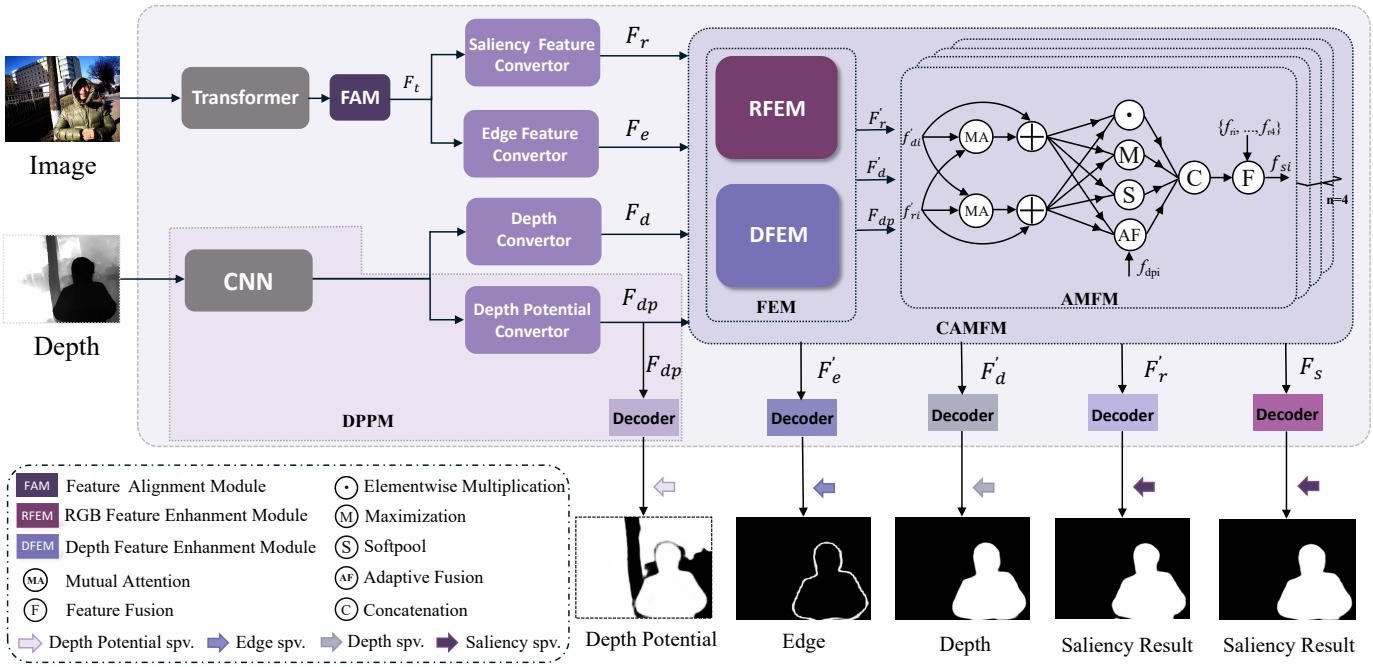

Fig. 3: The overall architecture of the proposed DPPNet. We first extract RGB and depth features by Transformer and CNN separately. Then, the RGB features are input FAM to obtain multi-level features. To achieve multi-task detection, the RGB and depth features are separately fed into two feature converters, extracting four independent multi-level features for saliency detection, edge detection, depth detection, and depth potential detection, respectively. Additionally, CAMFM is utilized to refine these features and carry out cross-modal fusion. Incorporating five decoders, we ultimately generate depth potential, edge, depth, saliency, and cross-modal fusion (saliency) maps. More details can be found in Sec. III-A.

Fourthly, slide the window, and then repeat the above steps two and three, until the completion of the whole picture content extraction.

$$i = i + S \tag{4}$$

$$j = j + S \tag{5}$$

where S is the step size during the sliding process and is set as 4.

After completing the above steps, we get a similarity matrix, and we multiply each element of this matrix by 255. Then, we resize the map to the same size as the original image by bilinear interpolation.

2) Generation of Depth Potential Maps: The depth potential maps are predicted by the Depth Pixel-wise Potential-aware Module (DPPM). DPPM is constructed with a predictor based on the ResNet-50 architecture, which has been trained using the aforementioned generated labels to assess the reliability of the depth map. The predictor produces a depth potential image, denoted as I_{dp}, which indicates the probability of the depth map being positive. A higher value in I_{dp} corresponds to a better quality depth region. Some predicted results are shown in the depth potential column of Fig. 1, 7. This module is trained concurrently with the network, without additional training. The training loss of DPPM adopts the pixel position-aware loss [50] to measure the structural similarity between the depth potential map I_{dp} and pseudo depth potential labels pGT.

C. Cross-modal Adaptive Multiple Fusion Module (CAMFM)

We propose a CAMFM to effectively fuse cross-modal features. The operation of this module can be described as (6).

$$F_s, F_r', F_e', F_d' = CAMFM(F_r, F_e, F_d, F_{dp}), \tag{6}$$

where F_s represents the cross-modal fusion result, utilized in deriving the final saliency map. F_r', F_e', F_d' are optimized saliency features, edge features, and depth features, respectively. F_r, F_e, F_d, and F_{dp} are the RGB features, edge features, depth features, and depth potential features, respectively. CAMFM consists of a feature enhancement module and an adaptive multiple fusion module. The following sections will provide a detailed introduction to these two modules.

1) Feature Enhancement Module (FEM): In cross-modal feature fusion, the characteristics of the features involved in the fusion process are the key factors influencing the fusion effectiveness. To avoid interference from noise and achieve effective fusion, we propose enhancing the RGB and depth features before performing cross-modal fusion. Regarding the inconsistency between RGB and depth features, we have adopted different enhancement strategies. These are described below:

RGB Feature Enhancement Module (RFEM). In order to capture sharper contour, we propose to incorporate an edge detection task to enhance the model performance. The edge features extracted by the edge detection task are used to enhance the structural information in the salient features.

Unlike normal one-way enhancement methods, we adopt a bidirectional iterative optimization scheme based on the characteristics between edge features and salient features. Different from [26], considering that the edge features contain more noise, we optimize the edge features with salient features first. The operation of this module can be described by (7).

$$F_r', F_e' = RFEM(F_r, F_e), \tag{7}$$

where F_r', F_e' represent the optimized feature output from RFEM. We use $F_r^n = \{f_{ri}^n, i = 1, 2, 3, 4\}$ and $F_e^n = \{f_{ei}^n, i = 1, 2, 3, 4\}$ to denote the features optimized for the *n*th time, where $n \in \{0, 1, 2, 3, 4\}$. f_{ri}^n denotes the *i*th level feature in F_r^n and the definition of f_{ei}^n is the same. When $n = 0$, F_r^0, F_e^0 represent the input RGB feature F_r and edge feature F_e. Accordingly, F_r^4, F_e^4 denote the feature optimized four times and are abbreviated as F_r', F_e'. In this module, we first optimize the edge feature F_e^0 with the salient feature F_r^0 to get F_e^1. Then we combine the salient feature F_r^0 with F_e^1 to get the optimized salient feature F_r^1 (see Fig.4). In order to obtain the best fusion effect, we repeat the two sets of optimized features more than 3 times.

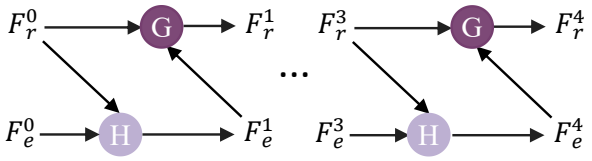

Fig. 4: The pipeline of RGB feature enhancement manipulation in RFEM. G and H represent the salient feature enhancement function and edge feature enhancement function, respectively.

The optimized process detail is as follows. Given the disparate nature of salient features and edge features, two different optimization algorithms are employed to enhance them individually. To achieve comprehensive optimization, we conduct optimization operations separately on features of different scales. The formulas for these operations are described as follows (8), (9).

$$f_{ei}^n = \begin{cases} H(f_{ei}^{n-1}, \{f_{ri}^{n-1}, .., f_{r4}^{n-1}\}), & i < 4, n = 1, 2, 3, 4 \\ H(f_{ei}^{n-1}, f_{ri}^{n-1}), & i = 4, n = 1, 2, 3, 4 \end{cases}, \tag{8}$$

$$f_{ri}^n = \begin{cases} G(f_{ri}^{n-1}, \{f_{ei}^n, .., f_{e4}^n\}), & i < 4, n = 1, 2, 3, 4 \\ G(f_{ri}^{n-1}, f_{ei}^n), & i = 4, n = 1, 2, 3, 4 \end{cases}, \tag{9}$$

where f_{ri}^{n-1} denotes the *i*th level salient feature in F_r^{n-1}. When n=1, f_{ri}^0 denotes the *i*th level salient feature of the input module. f_{ri}^1 denotes the *i*th level salient feature after one optimization. The meaning of f_{ei}^n is similar to f_{ri}^n. The functions $H(*)$ and $G(*)$ represent the edge feature enhancement operation and salient feature enhancement operation visible (10), (11) respectively.

$$H = f_{ei}^{n-1} + Conv(f_{ei}^{n-1} \odot \prod_{k=i}^{4} Up(f_{rk}^{n-1}))), \tag{10}$$

$$G = f_{ri}^{n-1} + Conv(Cat_{k=i}^{4}[f_{ri}^{n-1}, Up(f_{ek}^{n})]), \tag{11}$$

where $Conv$ stands for two 3 × 3 convolutional layers, $\odot$ means element-by -element multiplication, $\prod$ represents element-wise multiplied by part level salient features, $UP(*)$ for upsampling, $Cat[*]$ for concatenation.

Depth Feature Enhancement Module (DFEM). The actual value of depth in SOD lies in its geometric cues guiding foreground-background segmentation. However, the reliability of the depth map varies across scenes. To mitigate the negative impact of unreliable depth maps, we designed a depth enhancement manipulation. For promoting the model to selectively focus on relevant regions, we advocated leveraging the spatial attention mechanisms.

Specifically, we first employed the depth potential map to interact with the depth map to achieve selective enhancement and adopted the spatial attention mechanism to obtain more discriminative features (see Fig. 5). The operation of this module can be described by (12).

$$F_d' = DFEM(F_d, F_{dp}), \tag{12}$$

where $F_d' = \{f_{di}', i = 1, 2, 3, 4\}$ is the optimized depth feature. $F_d = \{f_{di}, i = 1, 2, 3, 4\}$, $F_{dp} = \{f_{dpi}, i = 1, 2, 3, 4\}$ are the depth and depth potential feature respectively.

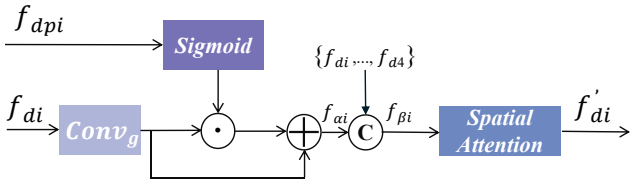

Fig. 5: The depth feature enhancement manipulation in DFEM.

To achieve better enhancement results, we optimize each level of depth features separately. The specific operation of DFEM is defined as (13) - (15).

$$f_{\alpha i} = Conv\big((Conv_g(f_{di})) \odot Sigmoid(Conv(f_{dpi}))\big) + Conv(Conv_g(f_{di})), \tag{13}$$

$$f_{\beta i} = \begin{cases} Conv(Cat[f_{\alpha i}, f_{di}, ..., f_{d4})]) & i < 4 \\ f_{\alpha i} & i = 4 \end{cases}, \tag{14}$$

$$f_{di}' = f_{\beta i} \odot Sigmoid(Conv(Cat[AvgPool(f_{\beta i}), MaxPool(f_{\beta i})])), \tag{15}$$

where $Sigmoid(*)$ is the sigmoid function, $MaxPool$, and $AvgPool$ denote the maximum and average pooling operations respectively, and $Conv_g$ is a convolution operation with a Gaussian kernel k and zero bias. The size and standard deviation of Gaussian kernel k are initialized with 5 and 0.5.

2) Adaptive Multiple Fusion Module (AMFM): The complementarity between RGB and depth maps is beneficial for promoting SOD. For example, RGB maps provide colour and texture, while depth maps provide spatial information, which is important for comprehensive scene understanding, accurate localization, and segmentation. To achieve effective cross-modality fusion, an adaptive multiple fusion strategy is employed across multiple scales of features to obtain more discriminative saliency features, as described in (16).

$$f_{si} = AMFM(\{f'_{ri}, ..., f'_{r4}\}, f'_{di}, f'_{dpi}), \tag{16}$$

where f_{si} represents the ith level feature in $F_s = \{f_{si}, i = 1, 2, 3, 4\}$. Similarly, f'_{ri}, f'_{di}, and f_{dpi} are the ith feature in F'_r, F'_d, and F_{dp}, respectively. Given the presence of four scale features, four AMFMs were employed to facilitate the fusion of cross-modal features at each level. The specific operations in AMFM are as follows:

Firstly, to exploit the correlations between the two modalities, we leverage a mutual attention operation [16] to get more discriminative features. Secondly, an adaptive multi-fusion strategy is adopted to realize cross-modality fusion. Unlike the general fusion scheme that treats different modalities equally and combines cross-modal features directly through addition or multiplication, we argue that multimodalities have different importance for the detection results and it is necessary to design an adaptive cross-modal fusion scheme to fully exploit their complementarity. To prevent contamination from unreliable depth maps, depth potential maps are used in the fusion process of RGB and depth maps. The process is as follows:

$$f_{ai} = f'_{ri} + f_{di} \odot Sigmoid(f_{dpi}) \quad i = 1, 2, 3, 4, \tag{17}$$

where $Sigmoid(*)$ is the sigmoid function that used to map real numbers to a range between 0 and 1.

Since the advantages of the different fusion strategies used for multimodality are different, we propose to combine some classical ones to make the most of their strengths. To obtain more discriminative features, we adopt maximization to select the maximum value between RGB and depth features, and softpool to retain the most relevant information. Besides, the classic operation element-wise multiplication is also used to capture the correlation between RGB and depth map. The operation can be defined as:

$$f_{maxi} = Max(f'_{ri}, f'_{di}) \quad i = 1, 2, 3, 4, \tag{18}$$

$$f_{pi} = SoftPool(f'_{ri}, f'_{di}) \quad i = 1, 2, 3, 4, \tag{19}$$

$$f_{mi} = f'_{ri} \odot f'_{di} \quad i = 1, 2, 3, 4, \tag{20}$$

where Max is maximization, $SoftPool$ denotes the operation similar to CIE [51] but we replace the max-pooling with a simplified soft pool to retain more relevant features. Finally, we concate the above result with the RGB features $F'_{ri} = \{f'_{ri}, ..., f'_{r4}\}$ to utilize the advantage of different fuse operations. The operation can be described as (21).

$$f_{si} = Conv(Cat[f_{ai}, f_{maxi}, f_{pi}, f_{mi}, F'_{ri}]) \quad i = 1, 2, 3, 4, \tag{21}$$

D. Loss Function

Our model is end-to-end trainable and is supervised by the salient ground truth mask G, salient edge ground truth mask G_e, and the pseudo depth potential ground truth mask pGT. S_r and S_d denote the prediction maps when using RGB and depth images, respectively. S_{sf} denotes the prediction map using their fusion representation. S_e represents the predicted edge map and S_{dp} denotes the predicted depth potential map. The overall loss function can be described as:

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}(S_{sf}, G) + \mathcal{L}(S_r, G) + \mathcal{L}(S_e, G_e) + \\ \mathcal{L}(S_{dp}, pGT) + \mathcal{L}(S_d, G),\end{aligned} \tag{22}$$

where $\mathcal{L}$ denotes the pixel position-aware loss [50]. It consists of the weighted binary cross entropy loss $\mathcal{L}_{wbce}$ and the weighted IoU loss $\mathcal{L}_{wiou}$, which is defined as (23).

$$\mathcal{L} = \mathcal{L}_{wbce} + \mathcal{L}_{wiou}, \tag{23}$$

$\mathcal{L}_{wbce}$ incorporates local structural information and allocates greater attention to challenging pixels than BCE. This approach may assist the model in emphasizing a broader receptive field. $\mathcal{L}_{wiou}$ assigns increased weights to hard pixels to underscore their significance. This strategy can additionally encourage the network to prioritize global structure. Overall, it integrates local structural information and global constraints, facilitating more effective guidance for network learning and the production of clear details. The definitions of $\mathcal{L}_{wbce}$ and $\mathcal{L}_{wiou}$ can be found in [50].

IV. Experiment

A. Implement Details

The proposed model is implemented using the PyTorch [52] library and trained on one NVIDIA Tesla V100 GPU with 32 GB memory. The backbone network consists of MAE [53] and Res2Net-50 [54]. Adam [55] algorithm is used to optimize the proposed model. The initial learning rate is set to $1e-4$ and reduced by a factor of 10 every 60 epochs. For training, the input resolutions of RGB and depth images are resized to 224 $\times$ 224. To improve the generalizability of the proposed learning algorithm, we adopt various data augmentation strategies, including rotating, random cropping, and random flipping. The model was trained over 140 epochs with a batch size of 30, taking around 8 hours. For testing, the RGB and depth images are resized to 224 $\times$ 224 before being fed into the model to obtain the predicted saliency map. The predicted saliency map is then resized to match the original size of the input images.

B. Datasets and Metrics

To validate the effectiveness of the proposed model, we evaluate our model on four widely used benchmark datasets, i.e. NJU2K [12] (1,985 samples), NLPR [56] (1,000 samples), STERE [57] (1,000 samples), SIP [9] (929 samples). For a fair comparison with SOTA methods, we adopt the same training set as in [34], which contains 1485 samples from NJU2K and 700 samples from NLPR for training. The remaining images

and all of the STERE and SIP datasets mentioned above will be used for testing.

We adopt some widely-used metrics to evaluate the model performance, including Mean Absolute error (M) [58], max F-measure (F_m) [17], S-measure (S_m) [59], and max E-measure (E_m) [60]. Mean Absolute error calculates the pixel-level average absolute error between the saliency maps and ground truth maps. F-measure is the weighted harmonic mean of precision and recall, used to comprehensively account for both the precision and recall rates. S-measure focuses on evaluating the structural similarity between object perception and regional perception. E-measure is an enhanced alignment metric that calculates both global and local similarity. Except for MAE, where a smaller value is preferable, higher values are better for the other metrics. Besides, we also draw the precision-recall (PR) and F-measure curves.

C. Ablation Studies

In this section, we explore the effect of different components in the proposed network via extensive ablation studies and comparisons. The contribution of each proposed block is presented in Tab. I. Additionally, a comparison of visualizations is provided in Fig. 6.

1) Effect of Depth pixel-wise Potential-aware Module(DPPM): In this subsection, we explore the effectiveness of DPPM in detecting the potential of depth maps. In the depth potential map, the black areas represent unreliable depth information, while the white areas are more reliable. Comparing the depth and depth potential maps in Fig. 6 and Fig. 7, our DPPM can estimate the confidence level of each point in the depth map. Even in an unreliable depth map, the DPPM can extract the valuable depth region as shown in Fig. 1. In addition, when comparing the GT and depth potential maps in Fig. 1, 6 and 7, it is evident that the predicted depth potential maps are similar to the salient ground truth. This is significant as it helps to focus on the effective depth region for better identification of the salient objects. The primary function of depth potential maps is to be used for depth map optimization, and its effectiveness will be demonstrated below.

2) Effect of Cross-modal Adaptive Multiple Fusion Module (CAMFM): To demonstrate the effectiveness of the proposed module, we conducted comprehensive ablation studies on each module and different combinations between them.

Effect of RGB Feature Enhancement Module (RFEM) Comparing No.1 and No.3 in Tab. I, the performance of the baseline model shows improvement after the integration of RFEM. Furthermore, we observed a significant decrease in performance when RFEM was excluded from the proposed model. The corresponding results are presented in the No.6 and No.8 of Tab. I. The visualization of the above comparison can be seen in Fig. 6. In contrast to the results obtained before

TABLE I: Quantitative evaluation for ablation studies of different modules of our model on four datasets. ↑ (↓) denotes that the higher (lower) is better. The best results are highlighted in red fonts.

No	DFEM(FEM)	RFEM(FEM)	AMFM	NLPR				NJU2K				STERE				SIP			
				M ↓	F_m ↑	S_m ↑	E_m ↑	M ↓	F_m ↑	S_m ↑	E_m ↑	M ↓	F_m ↑	S_m ↑	E_m ↑	M ↓	F_m ↑	S_m ↑	E_m ↑
1	✗	✗	✗	0.019	0.925	0.934	0.970	0.030	0.930	0.927	0.960	0.039	0.903	0.910	0.949	0.046	0.906	0.897	0.937
2	✓	✗	✗	0.018	0.929	0.938	0.970	0.028	0.934	0.931	0.964	0.033	0.918	0.920	**0.957**	0.042	0.912	0.898	0.938
3	✗	✓	✗	0.018	0.932	0.940	0.971	0.027	0.935	0.932	0.964	0.033	0.916	0.920	0.955	0.039	0.922	0.905	0.942
4	✗	✗	✓	0.018	0.935	0.940	0.974	0.030	0.932	0.928	0.961	0.041	0.899	0.905	0.943	0.049	0.897	0.887	0.927
5	✓	✓	✗	0.018	0.927	0.937	0.970	0.029	0.934	0.930	0.963	0.033	0.919	0.922	**0.957**	0.041	0.918	0.899	0.941
6	✓	✗	✓	0.018	0.930	0.937	0.972	0.030	0.930	0.927	0.959	0.034	0.918	0.920	0.955	0.039	0.922	0.906	**0.946**
7	✗	✓	✓	0.018	0.933	0.940	0.973	0.032	0.932	0.927	0.960	0.033	0.916	0.920	0.955	0.040	0.916	0.900	0.943
8	✓	✓	✓	**0.016**	**0.939**	**0.944**	**0.978**	**0.026**	**0.939**	**0.934**	**0.965**	**0.032**	**0.921**	**0.922**	0.956	**0.038**	**0.924**	**0.907**	**0.946**

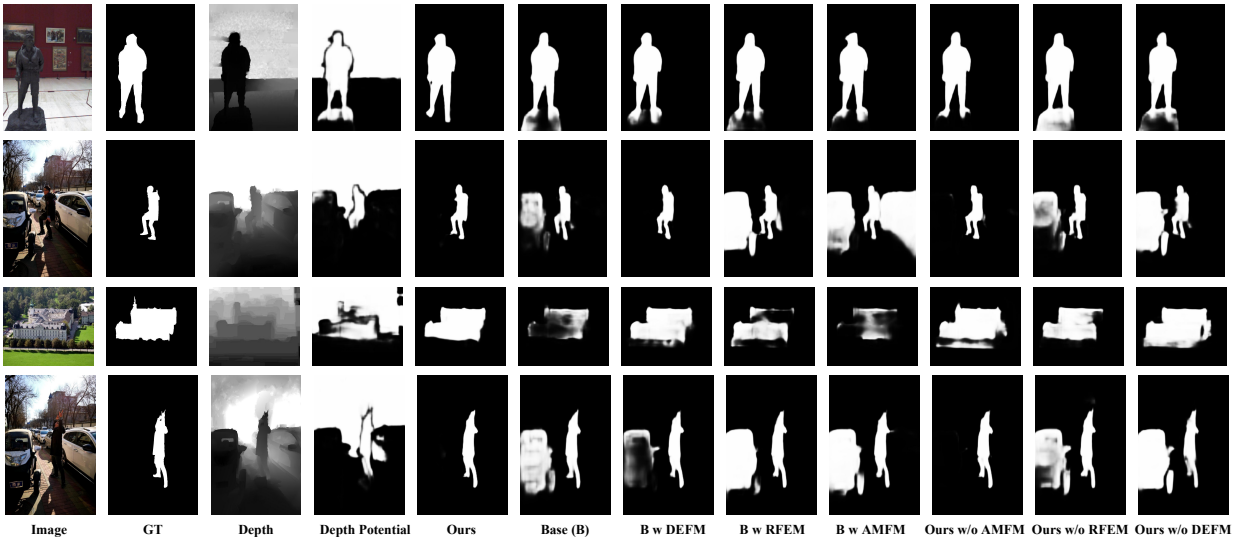

Fig. 6: Visual comparison with different modules. Base (B) indicates that no proposed module is used and multimodal features are fused by simple concatenation. 'w' and 'w/o' stand for with and without, respectively.

TABLE II: Ablation study conducted on the proposed multiple fusion scheme on four datasets. ↑ (↓) denotes that the higher (lower) is better. The best results are highlighted in red fonts.

No	Fusion Manipulation				NLPR				NJU2K				STERE				SIP			
	Multiplication	Maximization	Pooling	Adaptive Add	M ↓	F_m ↑	S_m ↑	E_m ↑	M ↓	F_m ↑	S_m ↑	E_m ↑	M ↓	F_m ↑	S_m ↑	E_m ↑	M ↓	F_m ↑	S_m ↑	E_m ↑
1	✗	✗	✗	✗	0.018	0.927	0.937	0.970	0.029	0.934	0.930	0.963	0.033	0.919	0.922	0.957	0.041	0.918	0.899	0.941
2	✓	✗	✗	✗	0.018	0.936	0.940	0.974	0.034	0.926	0.922	0.957	0.040	0.912	0.910	0.948	0.041	0.917	0.902	0.942
3	✗	✓	✗	✗	0.019	0.928	0.936	0.971	0.027	0.937	0.932	0.964	0.033	0.917	0.920	0.955	0.041	0.917	0.900	0.942
4	✗	✗	✓	✗	0.018	0.930	0.936	0.971	0.028	0.935	0.931	0.962	**0.032**	0.920	**0.923**	**0.958**	0.044	0.908	0.894	0.934
5	✗	✗	✗	✓	0.020	0.924	0.933	0.971	0.037	0.914	0.912	0.949	0.053	0.874	0.870	0.932	0.049	0.894	0.885	0.926
6	✓	✓	✗	✗	0.018	0.933	0.939	0.971	0.028	0.934	0.929	0.963	**0.032**	0.920	**0.923**	**0.958**	0.041	0.915	0.898	0.940
7	✓	✓	✓	✗	0.018	0.930	0.939	0.971	0.027	0.938	**0.934**	**0.965**	0.033	0.919	0.921	0.956	0.040	0.919	0.904	0.943
8	✓	✓	✓	✓	**0.016**	**0.939**	**0.944**	**0.978**	**0.026**	**0.939**	**0.934**	**0.965**	**0.032**	**0.921**	0.922	0.956	**0.038**	**0.924**	**0.907**	**0.946**

TABLE III: Quantitative comparison results with SOTA models on the publicly available RGB-D saliency detection datasets. ↑ (↓) denotes that the higher (lower) is better. The best results are highlighted in red fonts.

Methods	Published	NLPR				NJU2K				STERE				SIP			
		M ↓	F_m ↑	S_m ↑	E_m ↑	M ↓	F_m ↑	S_m ↑	E_m ↑	M ↓	F_m ↑	S_m ↑	E_m ↑	M ↓	F_m ↑	S_m ↑	E_m ↑
TriTransNet [61]	**ACM MM2021**	0.020	0.923	0.928	0.960	0.030	0.926	0.920	0.925	0.033	0.911	0.908	0.927	0.043	0.898	0.886	0.924
DCF [43]	**CVPR 2021**	0.021	0.918	0.924	0.957	0.035	0.893	0.890	0.924	0.039	0.911	0.902	0.927	0.052	0.899	0.876	0.920
DFM-Net [46]	**ACM MM 2021**	0.027	0.909	0.914	0.944	0.046	0.903	0.895	0.927	0.042	0.906	0.903	0.934	0.067	0.873	0.850	0.891
VST [29]	**ICCV 2021**	0.024	0.920	0.932	0.962	0.035	0.920	0.922	0.951	0.038	0.907	0.913	0.951	0.040	0.915	0.904	0.944
SPNet [16]	**ICCV 2021**	0.021	0.925	0.927	0.959	0.028	0.935	0.925	0.954	0.037	0.915	0.907	0.944	0.043	0.916	0.894	0.930
MobileSal [62]	**TPAMI 2021**	0.025	0.916	0.920	0.961	0.041	0.914	0.905	0.942	0.041	0.906	0.903	0.940	0.053	0.898	0.873	0.916
DCMF [63]	**TIP 2022**	0.027	0.915	0.921	0.943	0.044	0.908	0.903	0.929	0.041	0.909	0.907	0.931	0.067	0.873	0.853	0.893
MVSalNet [64]	**ECCV 2022**	0.022	0.931	0.930	0.960	0.036	0.923	0.912	0.944	0.036	0.921	0.913	0.944	-	-	-	-
SegMAR [65]	**CVPR 2022**	0.024	0.923	0.920	0.952	0.036	0.921	0.909	0.941	0.037	0.916	0.907	0.936	0.052	0.893	0.872	0.914
ZoomNet [66]	**CVPR 2022**	0.023	0.916	0.919	0.944	0.037	0.926	0.914	0.940	0.037	0.918	0.909	0.938	0.054	0.891	0.868	0.909
SPSN [15]	**ECCV 2022**	0.023	0.917	0.923	0.956	0.032	0.927	0.918	0.949	0.035	0.909	0.906	0.941	0.043	0.910	0.891	0.932
A2S [67]	**CVPR 2023**	0.034	0.862	-	0.931	0.060	0.862	-	0.908	-	-	-	-	0.051	0.873	-	0.925
HiDANet [14]	**TIP 2023**	0.021	0.929	0.930	0.961	0.029	**0.939**	0.926	0.954	0.035	0.921	0.911	0.946	0.044	0.919	0.892	0.927
PopNet [6]	**ICCV 2023**	0.022	0.925	0.926	0.956	0.031	0.931	0.920	0.949	0.032	**0.922**	0.916	0.947	0.046	0.911	0.885	0.926
CMIGNet [38]	**PR 2024**	0.020	0.924	0.933	0.965	0.029	0.931	0.928	0.959	**0.031**	0.919	0.920	**0.956**	-	-	-	-
DGFNet [47]	**TMM 2024**	0.021	0.902	0.928	-	0.032	0.914	0.921	-	0.035	0.896	0.911	-	0.048	0.879	0.883	-
TPCL [39]	**TMM 2024**	0.017	0.930	0.935	0.970	0.028	0.930	0.925	0.959	**0.031**	0.917	0.916	**0.956**	**0.038**	0.917	0.900	0.943
Ours		**0.016**	**0.939**	**0.944**	**0.978**	**0.026**	**0.939**	**0.934**	**0.965**	0.032	0.921	**0.922**	**0.956**	**0.038**	**0.924**	**0.907**	**0.946**

integrating RFEM, it is evident that refining the RGB features with RFEM has resulted in clearer contours in the detection results. This indicates that incorporating an edge detection task to optimize the RGB features is advantageous in enhancing the detection results.

Effect of Depth Feature Enhancement Module (DFEM) The advantages of refining the depth map to help the model focus on valuable depth regions and achieve a more accurate detection result can be seen by comparing the No.1 and No.2, as well as No. 4 and No.6 of Tab. I. After incorporating DFEM, the model's ability to locate saliency improved compared to the baseline model. After comparing the columns labeled "Base" and "Base w DEFM" in Fig. 6, it can be observed that when there is considerable interference in depth (i.e., when the depth map is unreliable), using DEFM helps the model achieve more accurate localization. To further validate the effectiveness of this module, we removed DEFM from the proposed model and found that unreliable information would contaminate the detection results, leading to segmentation anomalies (as observed in the comparison between the columns labeled "Our" and "Our w/o DEFM" in Fig. 6) and a decrease in model performance (as seen in the comparison between No.7 and No.8 in Tab. I). In the NJU2K, STERE, and SIP datasets, the quality of depth maps varies, with the SIP dataset having the highest depth quality and the STERE dataset having the lowest [43]. Comparing No.7 and No.8 in Tab. I, we observe that regardless of the reliability of the depth map, DFEM consistently helps the model achieve better performance.

Effect of Adaptive Multiple Fusion Module (AMFM) Comparing No.5 and No.8 in Tab. I, we find that model performance improves with multi-strategy fusion, with the most noticeable improvement observed on the SIP dataset. However, we also observe that multi-strategy fusion is not always effective. While leveraging the complementarity of multimodal inputs can enhance model performance, the quality of features themselves can affect fusion outcomes. Comparing No.1 and No.4 in Tab. I, we notice that if feature optimization is not performed before AMFM, the model's performance on the SIP dataset is inferior to simple fusion. We speculate that this may be due to noise in the features, which is amplified during multiple fusions, contaminating the detection results and leading to a decrease in model performance.

To gain further insights into the effectiveness of various fusion operations within multi-fusion strategies, we conducted ablation experiments targeting different fusion strategies. Comparing No.1 - No.5 in Tab. II, we observed that using only one fusion operation did not result in a significant performance improvement compared to simple concatenation. Comparing No.1, No.7, and No.8 in Tab. II, we found that by increasing the types and frequencies of fusion operations, there was a more noticeable improvement in model performance. This validates the effectiveness of the multi-fusion strategy.

D. Comparisons to SOTAs

We compare our model with 17 RGB-D SOD models including VST [29], SPNet [16], MobileSal [62], TriTransNet [61], DCF [43], DFM-Net [46], DCMF [63], SegMAR [65], ZoomNet [66], SPSN [15], MVSalNet [64], PopNet [6], A2S [67], HiDANet [14], CMIGNet [38], TPCL [39], and DGFNet [47].

1) Quantitative Comparisons: Fig. 8 shows the precision-recall (PR) curves and the F-measure curves for the three benchmark datasets. We can see that the solid red line which denotes the proposed method outperforms all other methods at most thresholds.

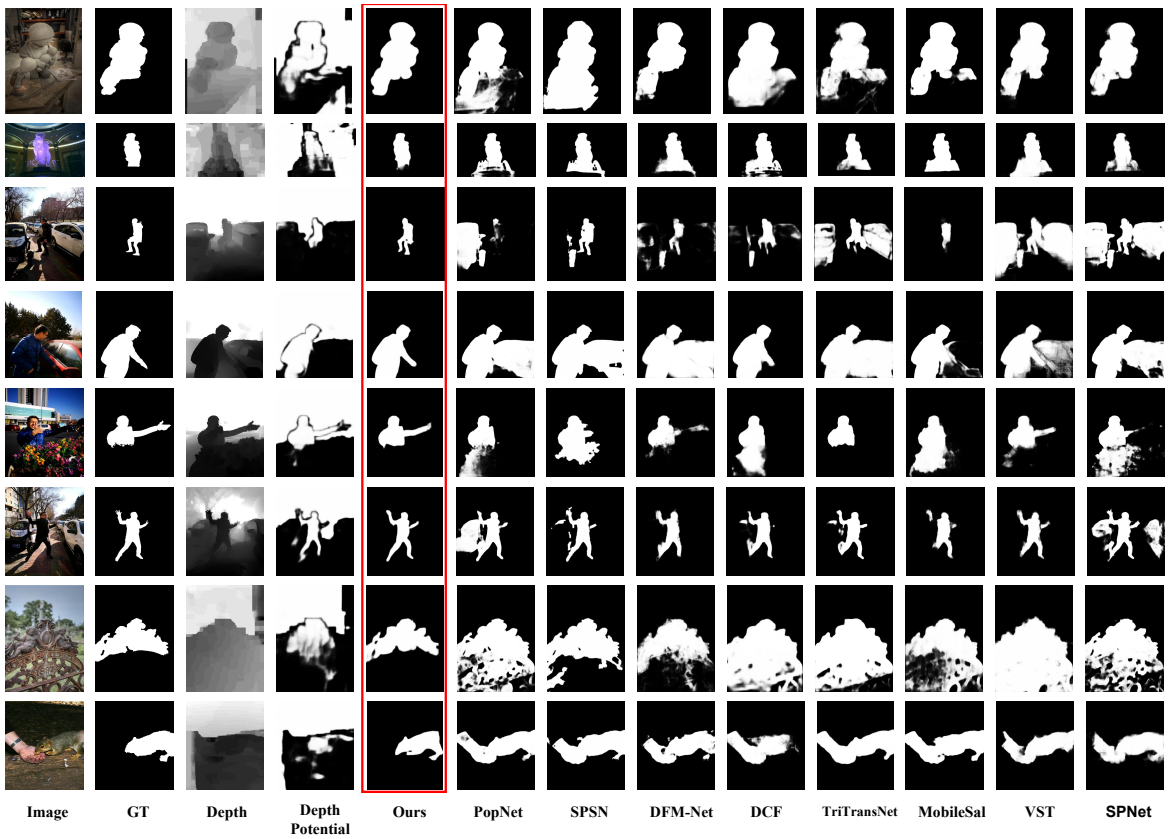

Fig. 7: Qualitative comparison of the proposed approach with SOTA methods. The detection results are presented for unreliable depth maps, categorized into two types: (a) low-quality depth maps (rows 1 to 6) and (b) extremely low-quality depth maps (last two rows).

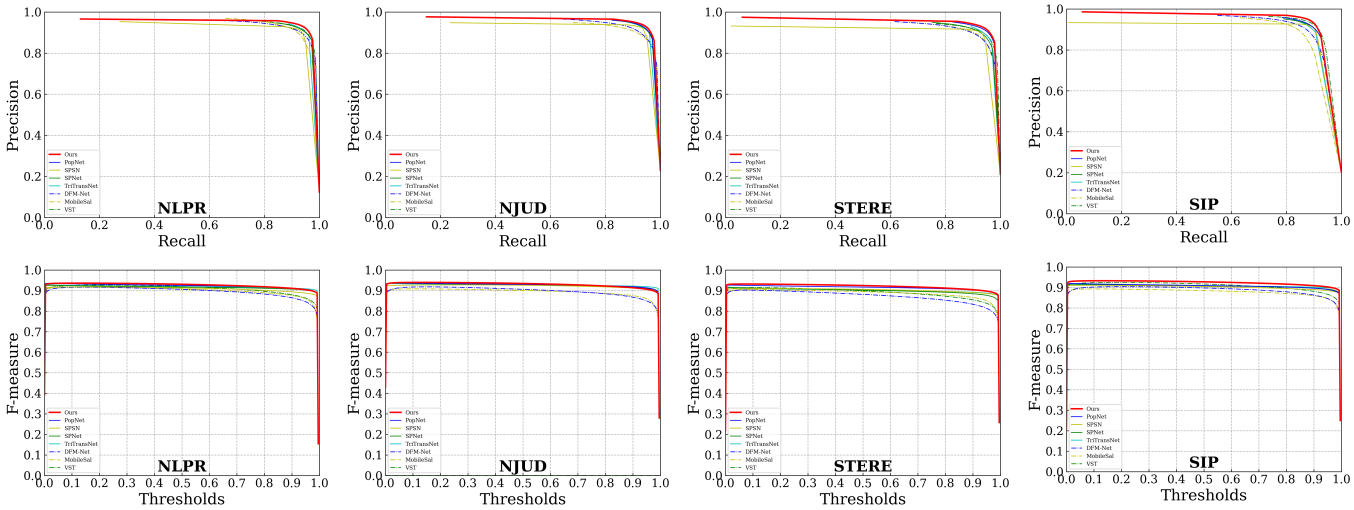

Fig. 8: Quantitative comparisons of the proposed approach and SOTA methods on four datasets. The first and second rows are the PR curves and F-measure curves of different methods respectively.

In addition, we evaluate and compare our proposed method with other mainstream methods in Tab. III. We use the Mean Absolute Error (M), max F-measure (F_m), S-measure (S_m), and max E-measure (E_m) as evaluation metrics. The results show that our model outperforms the majority of SOTA models on four RGB-D benchmark datasets across most evaluation metrics.

2) Visual Comparisons: In Fig. 7, we demonstrate the detection results of various methods under different low-quality depth maps, illustrating the effectiveness of our approach in extracting useful depth information even from low-quality depth maps. Note that all the saliency maps of the above

methods are pre-computed by the authors.

The quality of depth maps is classified into two categories, distinguished by the level of interference and the reliability of the object's depth information: low-quality and extremely low-quality. Low-quality depth maps are shown in rows 1 to 6 of Fig. 7, where there is considerable interference but the depth information of the detected objects remains significant. Comparing rows 1 to 6, we can see that the unreliable depth map causes serious contamination to the detection results of other methods, but with the help of the depth potential map, our method can still focus on the salient objects. The extremely low-quality quality depth maps are shown in rows 7 and 8 of Fig. 7, there is significant interference that obscures the depth information of the detected objects. Comparing rows 7 and 8, our model can still roughly locate the effective depth region and achieve more accurate detection even with these extremely low-quality depth maps. Besides, as shown in the 1st, 3rd, and 6th rows, it is worth noting that with RFEM, our model was able to not only highlight the salient region but also produce coherent edges.

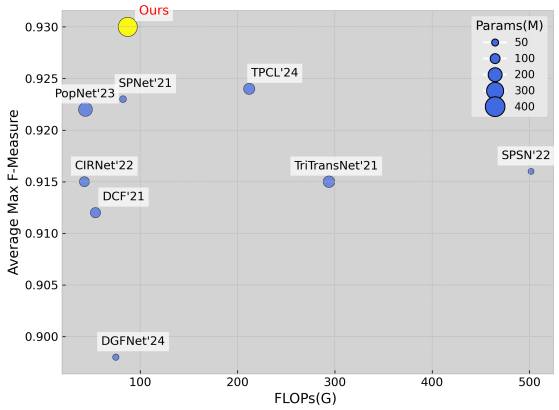

Fig. 9: Comparison of performance and complexity of different methods. The average maximum F-measure on four datasets [9], [12], [56], [57] and floating point operations (FLOPs) are plotted, with the circle size representing the number of model parameters (Params). Models located in the upper left corner indicate better performance, indicated by higher F-Measure and lower FLOPs.

3) Complexity Comparisons: Model complexity is a crucial performance metric, with Floating Point Operations (FLOPs) and parameters serving as common evaluation measures. To further explore performance, we use scatter plots to illustrate the relationship between model complexity and detection efficiency. A comparison of the overall performance of the proposed approach and SOTA methods on four datasets [9], [12], [56], [57] with model complexity is shown in Fig. 9. As illustrated, despite having a relatively large number of parameters, our model demonstrates superior performance compared to most mainstream methods with lower FLOPs on all four datasets.

E. Failure Cases & Disscusion

As shown in Fig. 10, in some scenarios our model will segment anomalies due to the insufficient use of the deep potential map or the incorrect prediction of the depth potential. The problem can be related to a variety of factors, such as the generation of pseudo-labels, the network structure, and the design of the cross-modal fusion scheme. In the next stage, we will address this problem by proposing some fusion schemes to fully exploit the deep potential map, improving the accuracy of the deep potential map generation, and designing some mechanisms to mitigate the negative impact of low-quality deep potential maps on the detection results.

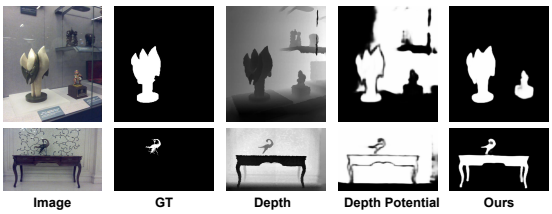

Fig. 10: Some failure cases of the proposed model.

V. Conclusion

We demonstrate a successful case of extracting valid information in unreliable depth maps to boost salient object detection performance. In this paper, a depth pixel-wise potential-aware module that predicts the confidence of each pixel in the depth map is proposed to effectively locate valuable information within the depth map, especially when the depth map is unreliable. A cross-modal adaptive multiply fusion module is designed to guide the model to better focus on effective information and prevent noise contamination. By incorporating various fusion strategies, the module maximizes the utilization of complementary multimodal data, thus achieving more accurate saliency localization. Experimental results show that our model outperforms other SOTA methods on multiple publicly available datasets. More importantly, our method provides a novel approach to dealing with low-quality depth maps, minimizing contamination and maximizing the extraction of valuable information. In our future work, we will focus on improving detection accuracy while reducing model size and increasing inference speed. Additionally, extending our approach to other multimodal dense prediction tasks and exploring its application on edge devices will be valuable areas for further research.

References

[1] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 1, pp. 20–33, 2017.

[2] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2020, pp. 263–273.

[3] P. Zhang, W. Liu, D. Wang, Y. Lei, H. Wang, and H. Lu, "Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps," *Pattern Recognition*, vol. 100, p. 107130, 2020.

[4] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8554–8564.

[5] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4819–4831, 2019.

[6] Z. Wu, D. P. Paudel, D.-P. Fan, J. Wang, S. Wang, C. Demonceaux, R. Timofte, and L. Van Gool, "Source-free depth for object pop-out," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1032–1042.

[7] A. Ciptadi, T. Hermans, J. M. Rehg *et al.*, "An in depth view of saliency." in *BMVC*, 2013, pp. 1–11.

[8] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "Rgbd salient object detection via deep fusion," *IEEE transactions on image processing*, vol. 26, no. 5, pp. 2274–2285, 2017.

[9] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on neural networks and learning systems*, vol. 32, no. 5, pp. 2075–2089, 2020.

[10] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 7012–7024, 2020.

[11] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 819–823, 2016.

[12] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 1115–1119.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[14] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, and C. Demonceaux, "Hidanet: Rgb-d salient object detection via hierarchical depth awareness," *IEEE Transactions on Image Processing*, vol. 32, pp. 2160–2173, 2023.

[15] M. Lee, C. Park, S. Cho, and S. Lee, "Spsn: Superpixel prototype sampling network for rgb-d salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 630–647.

[16] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving rgb-d saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4681–4691.

[17] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 248–255.

[18] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational visual media*, vol. 5, pp. 117–150, 2019.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[20] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 825–841.

[21] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3203–3212.

[22] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 1395–1403.

[23] K. R. Jerripothula, J. Cai, and J. Yuan, "Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2466–2477, 2018.

[24] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8779–8788.

[25] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7264–7273.

[26] J. Yuan, L. Xiao, K. Wattanachote, Q. Xu, X. Luo, and Y. Gong, "Fgnet: Fixation guidance network for salient object detection," *Neural Computing and Applications*, pp. 1–16, 2023.

[27] K. R. Jerripothula, P. Mukherjee, J. Cai, S. Lu, and J. Yuan, "Appfuse: An appearance fusion framework for saliency cues," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8261–8274, 2022.

[28] J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, "Poolnet+: Exploring the potential of pooling for salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 887–904, 2022.

[29] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4722–4732.

[30] J. Yuan, A. Zhu, Q. Xu, K. Wattanachote, and Y. Gong, "Ctif-net: A cnn-transformer iterative fusion network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[31] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Computational Visual Media*, vol. 7, pp. 37–69, 2021.

[32] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3051–3060.

[33] R. Cong, H. Liu, C. Zhang, W. Zhang, F. Zheng, R. Song, and S. Kwong, "Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 406–416.

[34] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7254–7263.

[35] F. Sun, P. Ren, B. Yin, F. Wang, and H. Li, "Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 2249–2262, 2024.

[36] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection," *Pattern Recognition*, vol. 86, pp. 376–385, 2019.

[37] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3052–3062.

[38] H. Zhu, J. Ni, X. Yang, and L. Zhang, "Cmignet: Cross-modal inverse guidance network for rgb-depth salient object detection," *Pattern Recognition*, p. 110693, 2024.

[39] J. Wu, F. Hao, W. Liang, and J. Xu, "Transformer fusion and pixel-level contrastive learning for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 1011–1026, 2024.

[40] J. Lin, L. Zhu, J. Shen, H. Fu, Q. Zhang, and L. Wang, "Vidsod-100: A new dataset and a baseline model for rgb-d video salient object detection," *International Journal of Computer Vision*, pp. 1–19, 2024.

[41] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3927–3936.

[42] F. Xiao, Z. Pu, J. Chen, and X. Gao, "Dgfnet: Depth-guided cross-modality fusion network for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 2648–2658, 2024.

[43] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu *et al.*, "Calibrated rgb-d salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9471–9481.

[44] X. Wang, S. Li, C. Chen, A. Hao, and H. Qin, "Depth quality-aware selective saliency fusion for rgb-d image salient object detection," *Neurocomputing*, vol. 432, pp. 44–56, 2021.

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

[46] W. Zhang, G.-P. Ji, Z. Wang, K. Fu, and Q. Zhao, "Depth quality-inspired feature manipulation for efficient rgb-d salient object detection," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 731–740.

[47] F. Xiao, Z. Pu, J. Chen, and X. Gao, "Dgfnet: Depth-guided cross-modality fusion network for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 2648–2658, 2024.

[48] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3907–3916.

[49] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[50] J. Wei, S. Wang, and Q. Huang, "F^{3}net: fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328.

[51] H. Bi, R. Wu, Z. Liu, H. Zhu, C. Zhang, and T.-Z. Xiang, "Cross-modal hierarchical interaction network for rgb-d salient object detection," *Pattern Recognition*, vol. 136, p. 109194, 2023.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[53] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 000–16 009.

[54] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[56] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 92–109.

[57] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 454–461.

[58] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[59] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 4548–4557.

[60] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.

[61] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4481–4490.

[62] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y.-C. Gu, and M.-M. Cheng, "Mobilesal: Extremely efficient rgb-d salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 10 261–10 269, 2021.

[63] F. Wang, J. Pan, S. Xu, and J. Tang, "Learning discriminative cross-modality features for rgb-d saliency detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 1285–1297, 2022.

[64] J. Zhou, L. Wang, H. Lu, K. Huang, X. Shi, and B. Liu, "Mvsalnet: Multi-view augmentation for rgb-d salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 270–287.

[65] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4713–4722.

[66] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2160–2170.

[67] H. Zhou, B. Qiao, L. Yang, J. Lai, and X. Xie, "Texture-guided saliency distilling for unsupervised salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7257–7267.



Junbin Yuan is pursuing the Ph.D. degree in software engineering from South China Normal University, Guangzhou, China. His research interests include computer vision and machine learning.



Yiqi Wang received the B.Eng. degree in network engineering from South China Normal University in 2023. Currently, she is pursuing the M.E. degree in Computer Technology at the same school. Her research interests include computer vision, especially image classification and semi-supervised learning.



Zhoutao Wang obtained the M.S. degrees in Software Engineering from the School of Computer at South China Normal University in 2019 and received the Ph.D. degrees in Computer Science and Technology from the School of Computer Science and Technology at Nanjing University of Aeronautics and Astronautics in 2024. His primary research focuses on deep learning models based on point cloud. Currently, his research focuses on leveraging both point cloud and image to enhance the accuracy and robustness of 3D object detection.



Qingzhen Xu received the Ph.D. degree in computer science from Sun Yat-sen University, in 2006. He is currently serving as a Professor with the School of Computer Science, South China Normal University. His research interests focus on computer vision and deep learning.



Bharadwaj Veeravalli (Senior Member, IEEE, IEEE-CS) obtained his Ph.D. degree from Indian Institute of Science in 1994. He is currently a tenured associate professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His current research interests include, Cloud/grid/cluster computing, Cybersecurity, analytics, and ML/DL applied to Computer Vision and image processing, and scheduling in parallel and distributed systems. He served as an associate editor for several computer society journals.



Xulei Yang (Senior Member, IEEE) received his PhD degree from Nanyang Technological University (NTU) in 2007. He is currently a principal scientist and group leader at Institute for Infocomm Research (I2R), A*STAR, previously the research head at YITU Technology Singapore, with more than 16 years of R&D experience in deep/machine learning for computer vision and healthcare. He has published more than 100 scientific papers and international patents in the fields of deep learning, 3D Vision and medical imaging.