Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 32(1): 643–669.
DOI: 10.3934/era.2024031
Received: 23 September 2023
Revised: 27 November 2023
Accepted: 10 December 2023
Published: 09 January 2024

Research article

Attention-guided cross-modal multiple feature aggregation network for RGB-D salient object detection

Bojian Chen*, Wenbin Wu, Zhezhou Li, Tengfei Han, Zhuolei Chen and Weihao Zhang

State Grid Fujian Electric Power Research Institute, No.64 Shoushan Road, Cangshan District, Fuzhou, China

* **Correspondence:** Email: 442953110@qq.com; Tel: +8618065072795; Fax:+8618065072795.

Abstract: The goal of RGB-D salient object detection is to aggregate the information of the two modalities of RGB and depth to accurately detect and segment salient objects. Existing RGB-D SOD models can extract the multilevel features of single modality well and can also integrate cross-modal features, but it can rarely handle both at the same time. To tap into and make the most of the correlations of intra- and inter-modality information, in this paper, we proposed an attention-guided cross-modal multi-feature aggregation network for RGB-D SOD. Our motivation was that both cross-modal feature fusion and multilevel feature fusion are crucial for RGB-D SOD task. The main innovation of this work lies in two points: One is the cross-modal pyramid feature interaction (CPFI) module that integrates multilevel features from both RGB and depth modalities in a bottom-up manner, and the other is cross-modal feature decoder (CMFD) that aggregates the fused features to generate the final saliency map. Extensive experiments on six benchmark datasets showed that the proposed attention-guided cross-modal multiple feature aggregation network (ACFPA-Net) achieved competitive performance over 15 state of the art (SOTA) RGB-D SOD methods, both qualitatively and quantitatively.

Keywords: salient object detection (SOD); RGB-D; feature aggregation; attention; cross-modal

1. Introduction

As a fundamental and essential vision task in computer vision, SOD aims to locate and identify the most visually eye-attracting objects in an image. With the continuous advance of SOD and deep neural network (DNN) technology, SOD has been widely applied in numerous computer vision-related applications, such as object classification and recognition [1, 2], target detection [3], semantic segmentation [4, 5], video object tracking [6], object discovery [7], image retrieval [8], simultaneous localization and mapping (SLAM) [9], style transfer [10], image translation [11] and image compression [12, 13].

Early SOD approaches [14–18] are mainly designed for RGB image and take it as input, which often suffers from performance degradation on more challenging cases, including similar texture and appearance of the foreground and background regions, occlusion, low-contrast light, complex and cluttered background etc. Thanks to the powerful representation ability of convolutional neural networks (CNN), the CNN-based RGB SOD methods [19–24] have achieved remarkable success compared to the traditional handcrafted feature-based RGB SOD methods. However, due to the loss of three-dimensional visual information in the two-dimensional RGB image, even CNN-based methods cannot completely solve the abovementioned issues, which still result in unsatisfactory performance when encountering complex and challenging scene images. Several surveys on RGB-based SOD [25–27] summarize its research progress in detail.

As is known to all, depth map contains spatial geometric and structure information and also provides indispensable complementary cues for RGB-based SOD and other vision tasks, including depth super-resolution [28], depth estimation [29, 30] etc. Due to the widespread popularity of smartphones and other advanced RGB-D sensors (e.g., Kinect), especially with the rise of portable depth cameras, depth maps are easy to obtain at a minimal cost. To this end, by combing the RGB iamges with auxiliary depth maps, recent research works [31–33] take both RGB and depth information as input and have verified its effectiveness in improving the detection and segmentation process. Compared with the RGB-based SOD method, the RGB-D SOD methods usually achieved promising performance in various challenging scenes.

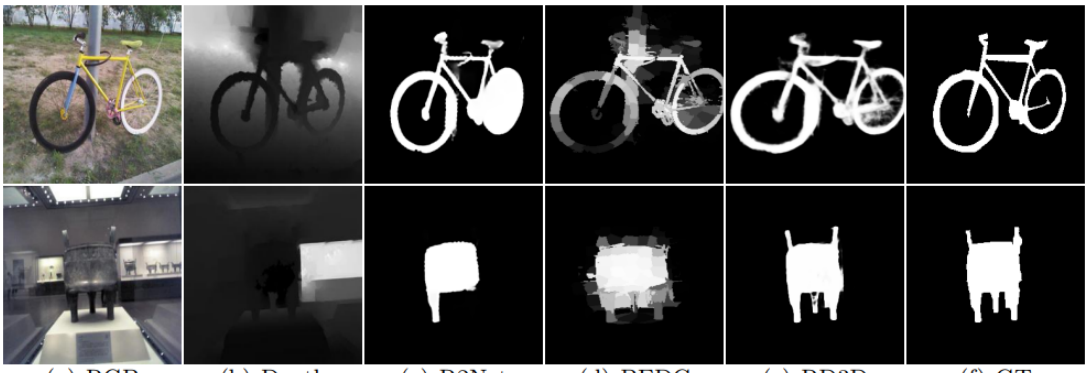

Figure 1. Two RGB-D input samples and the corresponding saliency maps generated by different types of SOD methods.(a) original RGB image, (b) depth map, (f) ground truth, (c),(d) and (e) are the saliency maps generated by DNN-based RGB SOD models, traditional RGB-D SOD methods and deep learning-based RGB-D SOD models.

As shown in Figure 1, although R2Net [34] is the latest SOTA RGB SOD method, it still has problems, such as incomplete detection and false detection when encountering some challenging scenarios. In addition, even if depth map is used, because the use of depth information is too simple and rough, the correlation between depth information and RGB information is not well mined and the traditional RGB-D SOD method BEDC [35] still has unsatisfactory detection performance. In contrast, the deep learning-based RGB-D SOD approaches RD3D [31] perform better. Therefore, although a number of previous RGB-D SOD methods have attempted to explore the effect and

contributions of depth map in SOD, the usage of auxiliary depth information brings several problems as follows. First, some low-quality depth maps can actually degrade and impair the detection performance of the RGB-D SOD model. Second, taking into account the feature complementarity of different modalities input and how to efficiently compute and represent depth-aware features for SOD task. RGB image provides rich semantic and appearance details, such as color and texture, but depth maps contain other useful supplementary information including shape, surface normals etc. How to use depth information correctly is an important issue that needs to be addressed. Third, how to effectively aggregate multilevel deep features from single modality data, and how to further integrate these cross-modal multilevel features from different modalities data, to suppress the background regions and completely highlight salient objects.

To address the aforementioned issues, we propose an attention-guided cross-modal feature pyramid aggregation network for RGB-D SOD, named ACFPA-Net. Specifically, to achieve cross-modal multilevel feature aggregation, we designed a novel cross-modal feature pyramid interaction module, denoted as CFPI. To address the performance degradation caused by low-quality depth maps, we developed a depth feature filtration module, denoted as DFF. Moreover, we also proposed a feature enhance and amplification module (FEA) to obtain more discriminative depth-aware features. Equipped with these useful modules, our ACFPA-Net can achieve better performance compared with multiple existing RGB-D SOD methods.

In summary, the major contributions of this paper are summarized as follows:

1) We propose an novel RGB-D SOD network via attention-guided cross-modal feature pyramid aggregation, named ACFPA-Net, which automatically extracts and aggregates the cross-modal multilevel visual information to highlight salient objects from various challenging scenes effectively.
2) By applying a residual convolutional block attention module (RCBAM), the CPFI module is designed to aggregate the cross-modal features for generating more discriminative depth-aware features.
3) We conduct extensive experiments on six RGB-D SOD benchmark datasets under four widely adopted evaluation metrics, which demonstrate the superiority and effectiveness of the proposed ACFPA-Net model against 15 recent SOTA models.

The rest of this paper is organized as follows. Section two first reviews related works of SOD. Then, section three describes the proposed attention-guided cross-modal feature pyramid aggregation network for RGB-D SOD. Next, the experimental results, corresponding disscussion and analysis are reported in section four. Finally, the concise conclusion of this paper is drawn in section five.

2. Related works

In the past two decades, SOD has been widely concerned by researchers due to its extensive applications, and many SOD methods have been presented. In this section, We briefly describe the most relevant works with SOD in terms of their input data. We first briefly review the related methods about the RGB SOD, then we introduce the development of RGB SOD methods and analyze the differences between these previously mentioned methods and our model. Among them, we focus on CNN-based RGB-D SOD approaches, as our proposed approach also falls into this category. In

addition to this, we also discuss the recent SOTA RGB-D SOD methods implemented with other deep neural networks instead of CNN.

2.1. RGB SOD

In the early days, the SOD methods were carried out on RGB images. These existing methods can be simply divided into two categories: Traditional RGB SOD methods and CNN-based RGB SOD models.

2.1.1. Traditional RGB SOD methods

The pioneering work [36] of SOD is proposed by Itti et al. This model extracts multiple feature maps including color, brightness, direction and gradient information to predict salient objects in RGB image. Inspired by this work, early traditional methods are mainly based on a wide variety of handcrafted features or intrinsic prior knowledge, such as local or global color contrast [17], center prior [37], spatial priors [16], objectness prior [38, 39], texture [40], background prior [14, 15], and etc., and have been proposed to identify and segment the salient objects. For example, Cheng et al. [17] proposed a regional contrast based SOD model, which simultaneously evaluates global contrast differences and spatial coherence. Contrary to [17], Yang et al. [14] ranked the similarity of the image pixels or regions with foreground cues or background cues via graph-based manifold ranking for saliency detection. There are too many such classic SOD methods to repeat here, so please refer to the review literatures [26, 41, 42]. In a nutshell, these methods are heavily depending on heuristic low-level features, thus lacking the guidance of high-level semantic cues. Although most of them are more computationally efficient, their performance can be severely degraded or even invalidated when challenging scenarios are encountered.

2.1.2. DNN-based RGB SOD models

Since 2012, deep learning has played an important role in various computer vision tasks. Naturally, deep learning based SOD models [25] have been explored. Benefited from the significant progress of CNN [43, 44], CNNs-based RGB SOD methods have quickly become the mainstream and have achieved impressive improvements compared with traditional SOD methods above. In 2015, several pioneering works [19–21] first introduced CNN into SOD. Among them, Zhao et al. [19] took global context and local context and proposed a multi-context deep learning framework for SOD. Wang et al. [21] proposed two deep neural networks DNN-L and DNN-G to learn local patch features for detecting local saliency and predicting the saliency score of object region based on the global features, respectively. Li et al. [20] integrated handcrafted low-level features with multiscale high-level semantic contrast feature extracted using CNN for saliency detection. These early models only take advantage top features of the backbone, which can hardly capture the detailed features of the salient objects because of operations such as downsampling.

Afterward, researchers focus on developing deep aggregation models, which aim to fuse multilevel features provided by the backbone. For instance, Liu et al. [45] leveraged the contextual attention module to extract local and global context cues. Chen et al. [46] proposed a reverse attention network to exploit the missing regions by erasing the predicted salient regions in side-output features. Liu et al. [47] designed a feature aggregation module to make the coarse-level semantic information well

fused with the fine-level features. Pang et al. [48] designed aggregation interaction modules to fuse multilevel features and propose self-interaction modules to get more efficient feature representations. In general, such approaches [49–51] are usually designed as an end-to-end encoder-decoder architecture with various effective feature manipulation strategies to extract, refine and integrate multi-scale multilevel features from CNN backbone network [44]. In other words, CNN is usually treated as the feature encoder (e.g., VGG [52], ResNet [44]) to generate multilevel visual features, and these extracted features are fed into a well-designed decoder for multi-scale feature fusion to produce the final saliency prediction results. It is worth mentioning that because of its excellent feature selection ability, the attention mechanism is very beneficial for SOD tasks. Therefore, some attention-based RGB SOD methods have also been explored and achieved superior performance [45, 53–55]. Liu et al. [45] utilized pixel-wise contextual attention to select global and local contextual information. Zhao et al. [54] proposed a pyramid feature attention network, which adopts channel-wise attention and spatial attention to focus more on valuable features.

Recently, many deep SOD models [50, 53, 56, 57] have been proposed to predict salient object contours and use it to enhance the object boundaries for SOD. For example, a novel boundary-aware model was presented in BASNet [50] to enhance the boundaries of salient objects by incorporating a boundary localization stream. Zhao et al. proposed an edge guidance network (EGNet) [57] to explicitly model complementary salient object information and salient edge information within the network to preserve the salient object boundaries. Detailed introductions of SOD works can refer to recent popular survey [25]. Although the RGB SOD methods perform well, when faced with challenging scenarios such as low illumination and contrast, transparent objects and complex and cluttered scenes, these methods are still a bit overwhelming. This is mainly because RGB images contain only visual appearance information and no rich spatial geometry and structural information. In contrast, depth maps can provide rich spatial geometric cues that RGB images cannot provide and are helpful for detecting salient objects. This type of RGB-D SOD approach is discussed in the next section.

2.2. RGB-D SOD

Over the past decade, by leveraging depth information, a large number of RGB-D SOD methods have been proposed. In this section, similar to RGB SOD, we roughly classify previous RGB-D SOD approaches into two categories: Traditional RGB-D SOD methods, and CNN-based RGB-D SOD models. Especially in recent years, the latter have become mainstream and have achieved encouraging progress compared with the former.

2.2.1. Traditional RGB-D SOD methods

Similar to traditional RGB SOD methods, early RGB-D SOD methods extract handcrafted features (e.g., local and global contrast [58–60], spatial prior [60, 61], background prior [62] etc.) from image pairs (RGB images and depth maps) and fuse them for detecting salient objects. As a matter of fact, depth cue is often treated as a complementary prior together with other specific prior information from RGB images to assist saliency detection on RGB-D images. The pioneering RGB-D SOD study [58] computed the global disparity contrast and domain knowledge of stereoscopic images to measure stereo saliency, and built a stereo saliency analysis benchmark dataset STEREO. Subsequently, Cheng et

al. [60] measured salient value using spatial bias and contrast cues (color contrast and depth contrast), and built the DES dataset for RGB-D SOD. In the same year, Peng et al. [59] proposed a multi-contextual contrasted-based saliency detection method and built the NLPR dataset. The latter two datasets, DES and NLPR, directly provide depth maps collected by the Kinect device. The emergence of these datasets largely stimulates the study in RGB-D SOD. For example, Ren et al. [62] integrated region contrast with the background prior, depth and surface orientation prior to generate a coarse saliency map, and reconstructed the final saliency map by a saliency restoration stage. Wang et al. [63] developed a multi-stage SOD framework via minimum barrier distance transform and multilayer cellular automata-based fusion with 3-D spatial prior and depth bias. Cong et al. [64] proposed a depth-guided transformation and optimization model to incorporate depth map into the existing RGB SOD methods for boosting the performance of RGB-D SOD. Although these traditional methods have achieved promising performance, the low-quality depth maps and the handcrafted features limit their generalization and performance improvement in complex scenarios.

2.2.2. Deep learning-based RGB-D SOD models

Recently benefitting from the depth cue, which contains rich spatial structural information, deep learning-based RGB-D SOD methods have achieved significant progress. Among the first such studies, Qu et al. [65] designed a simple CNN-based model to learn the interaction mechanism of RGB and depth-induced saliency features for RGB-D SOD. Piao et al. [66] designed a depth refinement block to extract and fuse multilevel paired features and combined depth cues with multi-scale context features for locating salient objects. Although its model architecture is relatively simple and straightforward, its detection performance has improved significantly over traditional methods. CNN-based RGB-D SOD methods [67, 68] can adaptively extract and fuse discriminative complementary features from RGB-D image pairs. Among them, fusion-based approaches have devoted significantly to RGB-D SOD and have achieved excellent performance. In terms of the fusion strategy of RGB and depth modal information in most previous papers [69–71], RGB-D SOD methods can be roughly divided into the following three popular categories: Data-level fusion, feature-level fusion, and result-level fusion, especially the intermediate one.

Data-level fusion. This type of fusion strategy directly concatenates three-channel RGB image and single-channel depth map together as four-channel input image. Models that employ this fusion strategy usually train a single-stream SOD model with the composited four-channel input. Typically, Wang et al. [72] proposed a data-level recombination strategy to fuse RGB with depth data, and applied a lightweight designed triple-stream network to predict salient objects. To explore the effect of the depth map, DANet [73] directly uses the depth map to guide the early fusion between RGB and depth modality. Subsequently, a joint learning and cooperative fusion model is proposed in [74] to exploit the cross-modality complementary information. With shared parameters for feature extraction, these methods largely reduce the number of parameters while tending to learn compromising features between modalities.

Result-level fusion. This type of approach commonly adopts two-stream CNNs on the RGB image and the depth map, respectively, to obtain two initial RGB-related and depth-related saliency prediction maps, and then fuses them in a variety of ways to generate the final saliency map, such as concatenation, addition, multiplication etc. For typical examples, Han et al. [75] adopted a transfer learning strategy to integrate the feature representations of RGB and depth to generate the final

saliency map. In AFNet [76], a two-stream CNN was designed to extract features and predict saliency map from RGB and depth modality respectively, and used a saliency fusion module to learn a switch map to fuse the predicted saliency maps. Li et al. [77] proposed an information conversion network (ICNet) with encoder-decoder architecture for RGB-D SOD, in which an information conversion module was used to fuse high-level RGB and depth features in an interactive and adaptive way. Since only feature interactions are performed on the prediction maps, such methods cannot fully characterize the correlations between the two modalities.

Feature-level fusion. Such fusion methods generally adopt a two-stream network structure to extract multi-scale RGB and depth features separately, and then aggregate cross-modality features at multiple levels by a specially designed cross-modal fusion unit to generate the final saliency prediction map. To better explore the complementary values from each other, both [78] and [79] design a complementary-aware fusion module to select and combine multi-modal features. However, the low-quality depth maps may lead to poor fusion results. DQSD [80] integrates a novel depth quality-aware subnet to assess the depth quality before conducting the selective RGB-D fusion. Fan et al. [81] proposed a bifurcated backbone strategy to split the multilevel features into teacher and student features. Ji et al. [66] proposed a depth-induced multi-scale recurrent attention network to learn the internal semantic relation of the fused features and optimize local details with memory-oriented scene understanding. Jin et al. [82] supplemented the depth features with a depth map estimated from RGB images and fused the bimodal features in two stages according to the hierarchy. Wu et al. [70] designed an implicit depth restoration strategy to enhance the learning of features by the backbone network during the training phase. Different attention-based mechanisms are introduced in [68, 83, 84] to explore complementary cross-modality information for improving the performance.

In general, data-level and result-level fusion strategies are more efficient, and feature-level fusion is more accurate. The proposed model in this paper belongs to feature-level fusion based ones. In fact, there are some inspiring related models [85–87] based on other neural networks that perform well. Discussing these models in detail is beyond the scope of this paper, so please refer to the recent survey [27] for more details. Although great performance improvements have been made, existing models do not deal with the intra and inter-feature interaction issues well, as most of them only regard depth map as the supplement of RGB image and ignore the correlations between the two, and still encounter problems such as incomplete feature aggregation and partial boundary loss. For feature fusion itself, in addition to cross-modal feature fusion, cross-level feature fusion and cross-scale feature fusion are also crucial for RGB-D SOD methods. In this context, low-level detail features and high-level semantic cues are fused progressively, and multi-scale features are aggregated to complement contextual information at different levels of detail. Although some methods have explored cross-modal and cross-scale feature fusion, they have not considered the importance of multi-scale features [88–90]. Different from these aforementioned methods, in order to balance high accuracy and efficiency, in this paper, we design a cross-modal pyramid feature interaction module to integrate features from RGB and depth modalities under the premise of fully exploiting multi-scale information, and introduce a cross-modal feature decoder to aggregate these fused features to generate the final prediction results.

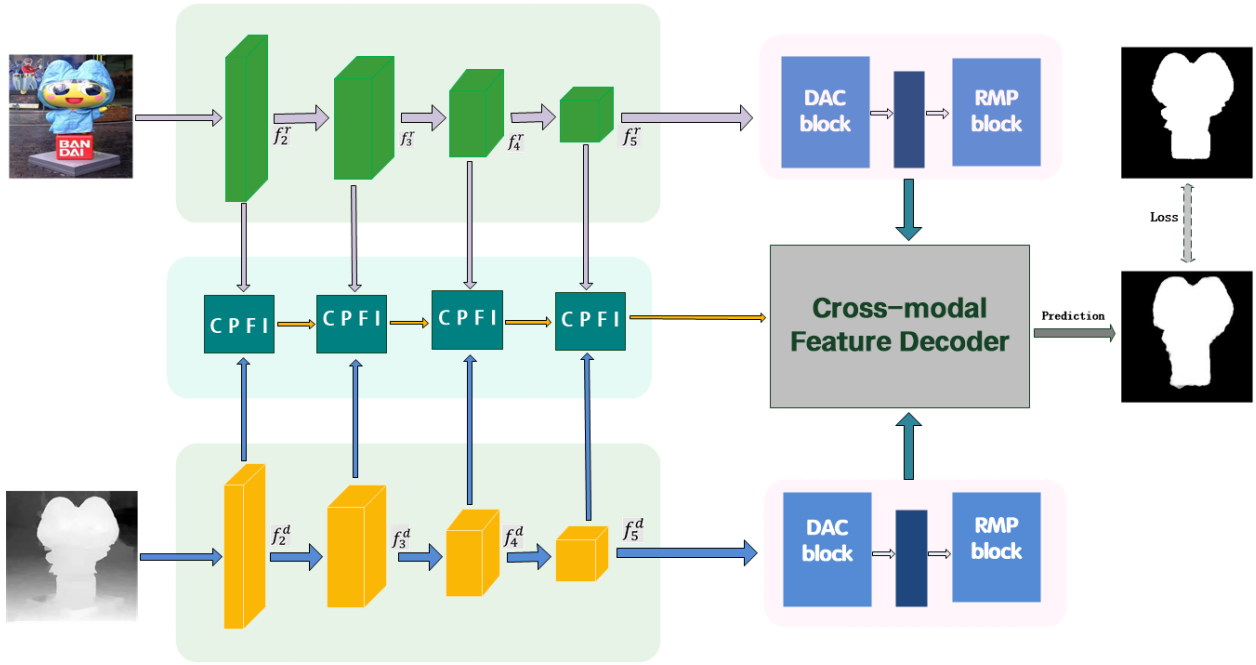

Figure 2. The overall architecture of the proposed ACFPA-Net. It mainly consists of a backbone network, CPFI module and cross-modal multilevel feature decoder (CMFD) module.

3. Proposed methods

We first briefly describe the overall backbone network architecture of the proposed ACFPA-Net in in Section 3.1. Second, we give a detailed explanation for the CPFI module in Section 3.2, then the CMFD and pyramid dilated convolution module are illustrated in Sections 3.3 and 3.4, respectively. Finally, the stage-wise intermediate supervision is introduced in Section 3.5.

3.1. Backbone network

The overall framework of our proposed ACFPA-Net is shown in Figure 2, which follows a standard encoder-decoder architecture. The feature encoder contains an RGB and depth stream backbone networks for separate feature extraction, which is constructed based on a ResNet-like pretrained model that removes the last fully-connected layer and the global average pooling layer. First, a pair of RGB and depth images are fed into two dual-stream feature encoders for multilevel feature extraction, then the extracted features are progressively integrated and refined by multiple cascading CPFI modules. Subsequently, the fused features from the CPFI module, along with high-level single-modal semantic features from both RGB and depth stream, are further fed into the feature decoder for cross-modal fusion. In order to expand the receptive field to obtain richer high-level semantic features, it should be noted that single-modal features need to be enhanced by a pyramid dilated convolution (PDC) module [91] before being fed into the feature decoder. These modules are elaborated in detail in subsequent sections.

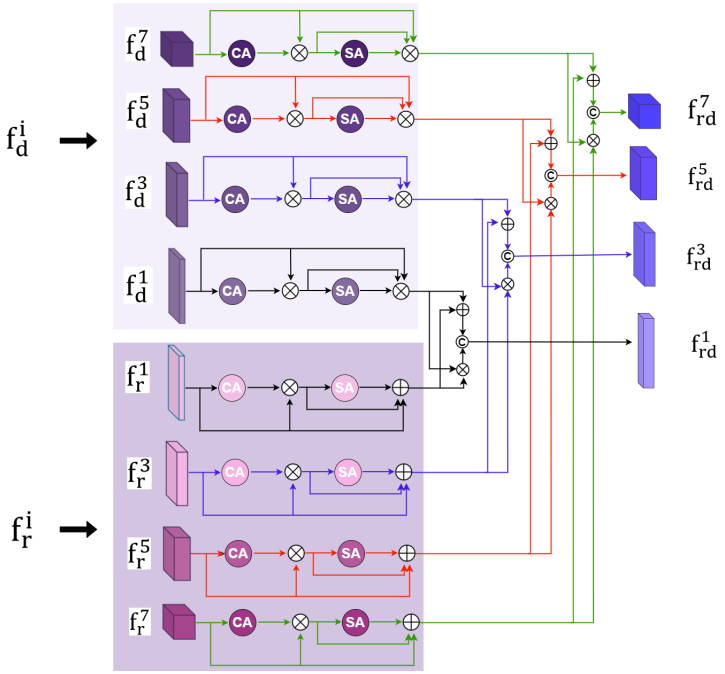

Figure 3. Illustration of the CPFI module.

3.2. CPFI module

The CPFI module is designed to integrate features from RGB and depth modalities, as shown in Figure 3, and it is divided into two parts: Single-modal feature enhancement and multi-modal multilevel feature interaction. In the single-modal feature enhancement, feature maps f_r^i and f_d^i , $i = 1, 3, 5, 7$ with the same scale, along with shallow features, are taken as inputs. First, the single-modal features are expanded using dilated convolutions to obtain rich features with different levels of receptive fields, then the RCBAM is a modification from CBAM [92], which adds an additional residual connection to better refine the obtained features. This process can be described in detail as follows:

$$rcbam() = SA(CA(f) \otimes f) \otimes CA(f) + f, \tag{3.1}$$

where SA and CA represent spatial attention, and channel attention, respectively, f denotes the input feature, and $\otimes$ is the multiplication operation.

Specifically, linear operation is employed to interact with the enhanced single-modal features. We first perform feature aggregation on single-modal features with the same dilation rate, obtaining rich fused features with different receptive fields. Then, we concatenate these features and utilize the previous layer's feature containing abundant detailed information as a supplement to obtain the final fused feature, as follows:

$$f_{rd}^{i(\lambda)} = cat(rcbam(f_r^{i(\lambda)}) + rcbam(f_d^{i(\lambda)}), rcbam(f_r^{i(\lambda)}) \otimes rcbam(f_d^{i(\lambda)})) \tag{3.2}$$

$$f_{cpfim}^{i} = \begin{cases} cat(f_{rd}^{i(1)}, f_{rd}^{i(3)}, f_{rd}^{i(5)}, f_{rd}^{i(7)}) & i = 2 \\ cat(f_{rd}^{i(1)}, f_{rd}^{i(3)}, f_{rd}^{i(5)}, f_{rd}^{i(7)}) + f_{cpfim}^{i-1} & i = 3, 4, 5, \end{cases} \tag{3.3}$$

where cat stands for the concatenate operation and $f^{i(\lambda)}$ represents the feature from the i-th layer with a dilation rate of $\lambda(\lambda = 1, 3, 5, 7)$.

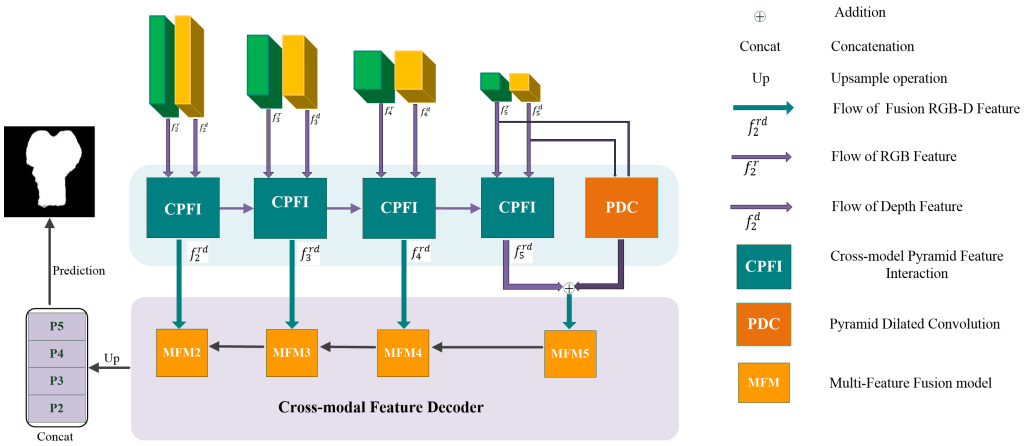

Figure 4. Illustration of the CMFD module.

3.3. CMFD

Existing research has shown that features at different levels in a network can reflect different characteristics of objects. Specifically, shallow features can provide richer local information, while deep features contain semantic-level global information. In the encoding stage, we obtain high-level features with abundant information. Through experiments, we found that effectively utilizing shallow features in the decoding stage can significantly improve the model's performance. To achieve this, we introduce the mutually guided cross-level decoder [93] and modify it as CMFD module to aggregate multi-scale and comprehensive features from both top-down and bottom-up pathways. Specifically, as shown in Figure 4, the output features of the CPFI module are first aggregated with global semantic guidance features f_{pdcd} and f_{pdcr}; then the high-level semantic aggregated features $\left\{f^j_{fusion}\right\}^5_{j=2}$ are fed into the feature fusion (FF) module in the top-down pathway to interact with shallow information and complement the detailed features.

To fully explore the complementary nature of local and global information, the multi-scale feature fusion module (MFM) integrates outputs from the FF side $\left\{f^j_{ff}\right\}^5_{j=2}$ in the bottom-up pathway and high-level semantic aggregated features $\left\{f^j_{fusion}\right\}^5_{j=2}$. The bidirectional pathway aggregates multi-scale cross-modal fusion features from low and high layers, enabling the prediction of complete structures and clear boundaries. Figure 4 illustrates the implementation details of this module. In each FF module, the input consists of high-level semantic aggregated features $\left\{f^j_{fusion}\right\}^5_{j=2}$ and FF features $\left\{f^j_{ff}\right\}^4_{j=2}$ from the upper layer. Therefore, the feature fusion in FF can be described as:

$$f_i = (\widetilde{w}_i * f^i_{fusion}) \otimes (\widetilde{w_{j+1}} * AP(F_j + 1)) \tag{3.4}$$

$$\begin{cases} F_j = w_i * f^i_{fusion} & i, j = 5 \\ F_j = w_i * f_i \oplus w_{j+1} * U(F_{j+1}) & i, j \in \{2, 3, 4\}, \end{cases} \tag{3.5}$$

where $\oplus$ and $\otimes$ represent element-wise addition and multiplication and U is the bilinear interpolation upsampling operation. AP denotes the global average pooling operation. $\{\widetilde{w}_i\}_{j=2}^{4}$ and $\left\{\widetilde{w_{j+1}}\right\}_{j=2}^{4}$ represent 3×3 convolution operations with PReLU activation and 1×1 convolution operations with ReLU activation, respectively, both of which transform the channels of $\left\{f_{fusion}^{j}\right\}_{j=2}^{5}$ into 256. $\{w_i\}_{i=2}^{5}$ and $\left\{w_{j+1}\right\}_{j=2}^{4}$ denote 3×3 convolution operations without activation. F_j is the output of the FF module and is simultaneously passed to the next layer's FF module as input.

The MFM will bottom-up receive the outputs $\left\{F_j\right\}_{j=2}^{5}$ from FF and cross-modal fusion features $\left\{f_{fusion}^{j}\right\}_{j=2}^{5}$ to generate multi-scale prediction results:

$$\begin{cases} S_k = scSE(\widetilde{w}_{\sum}^{k} * (F_j \oplus (\widetilde{w}_{sk}^{i} * f_{fusion}^{i})) \oplus (\widetilde{w}_{\sum}^{k-1} * S_{k-1}^{i}))) & i, j, k \in \{3, 4, 5\}, \\ S_k = scSE(\widetilde{w}_{\sum}^{k} * (F_j \oplus (\widetilde{w}_{sk}^{i} * f_{fusion}^{i}))) & i, j, k = 2, \end{cases} \tag{3.6}$$

where $\oplus$ denotes element-wise addition. $\left\{w_{sk}^{i}\right\}_{i=2}^{5}$ and $\left\{\widetilde{w}_{\sum}^{k}\right\}_{k=2}^{5}$ represent 1×1 convolution operations with ReLU activation. $\left\{\widetilde{w}_{\sum}^{k-1}\right\}_{k=3}^{5}$ denotes downsampling by convolution operations. The MF module is used to filter out unnecessary information and obtain multilevel prediction outputs $\{S_k\}_{k=2}^{5}$. These multilevel predictions are concatenated after upsampling and then passed through a convolutional layer to obtain the final saliency map S_{map}:

$$S_{map} = P(cat_{k=2}^{5} U(S_k)), \tag{3.7}$$

where U is the upsampling operation, and P means the 3×3 prediction convolution layer.

3.4. PDC module

The PDC takes high-level features f_{rgb}^{5} and f_{d}^{5} as input to capture rich semantic and positional information, providing global semantic guidance in the decoding stage. In real-world scenarios, salient objects may vary in size. To address objects of different scales, the PDC module includes two parts: Dense atrous convolution (DAC) unit and residual multikernel pooling (RMP) unit. The DAC unit takes inspiration from the structure of Inception-ResNetV2 and consists of four parallel branches, by replacing convolutions at different scales with dilated convolutions. Specifically, in DAC, different numbers and dilation rates of atrous convolutions are stacked in the four branches to expand the receptive fields. We use atrous convolutions with dilation rates of one, three, and five, resulting in receptive fields of three, seven, and nine for each branch. Additionally, each branch undergoes a 1×1 convolution for rectified linear activation. Finally, similar to ResNet's skip connections, the aggregated features from different receptive fields are added to the original features to obtain the richer fused features. Inspired by spatial pyramid pooling (SPP) [94], the RMP unit further encodes multi-scale contextual features of objects extracted from the DAC unit without requiring additional learning weights. In this work, we set four receptive field scales: 2×2, 3×3, 5×5 and 6×6, resulting in four feature maps of different sizes. After each pooling level, a 1×1 convolution is used to reduce the channels of the feature maps by 256 for reducing computational cost, then we upsample the low-dimensional features to the same size as the original features by using bilinear interpolation. Finally, we concatenate the original features with the upsampled features as the output feature.

3.5. Stage-wise intermediate supervision

To facilitate the training of the network, we apply feature supervision at multiple stages. Specifically, in addition to the predicted maps from the network's output, we perform 3 × 3 convolutions on the feature maps obtained from the CPIF and PDC modules to obtain corresponding predicted maps. All the prediction results are then adjusted to the same resolution as the input image through bilinear interpolation.

We use weighted cross-entropy and weighted intersection over union as the loss functions. Therefore, the final loss of the network is defined as follows:

$$l(f) = \ell_{wbce}(f, G) + \ell_{wiou}(f, G), \tag{3.8}$$

$$Loss = \lambda_1 l(S_{map}) + \lambda_2 \sum_{i=2}^{4} l(f^i_{cpfim}) + \lambda_3 (l(f_{pdcr}) + l(f_{pdcd})), \tag{3.9}$$

where G represents the ground truth, $\{\lambda_i\}_{i=1}^{3}$ is a hyperparameter, f_{pdcd} and f_{pdcr} are the output features from the PDC module, and S_map is the final predicted map of the network.

4. Experimental results and discussion

We first described the implementation details of the proposed ACFPA-Net model in Section 4.1, then introduced the six RGB-D SOD benchmark datasets and five commonly used evaluation metrics in Section 4.2 and in Section 4.3. After that, the comparisons with 15 SOTA CNN-based methods are conducted. Finally, we conduct a series of ablation studies to validate the effectiveness of our proposed modules. We conducted a quantitative and qualitative comparative analysis with 16 SOTA RGB-D SOD models in Section 4.4 and all-round ablation studies in Section 4.5.

4.1. Implementation details

The proposed ACFPA-Net is implemented based on pytorch and trained with a single NVIDIA GeForce RTX 3090 GPU, and the total training time in total takes five hours corresponding to 70 epochs. Our model architecture is independent of the backbone network. During the training phase, for the sake of fairness, we adopt the ResNet-50 [44] as the backbone network for both RGB and depth streams, which is initialized by the pretrained parameters on ImageNet. For the depth stream, we adopt gray color mapping to transform the single-channel depth map into a three-channel image as input. Several common transformations including random flipping, rotating and cropping are adopted for data augmentation to prevent model overfitting. Multi-scale training is also applied; that is, all the training input samples are resized into [320, 352, 384]. We set the maximum epoch and batch size as 70 and 10, respectively. The AdamW optimizer [95] is employed for optimizing the proposed network model. The corresponding learning rate is initially set to 1e-5 and dynamically adjusted every 20 epochs with weight decay 0.1. During inference stage, all the test images are uniformly fixed to 352 × 352 resolution, then fed into the network to generate the final saliency prediction map without any other post-processing steps. The inference time of our method is about 0.06 second for an image.

4.2. Benchmark datasets

Training dataset. To verify the effectiveness and generalization ability of the proposed model ACFPA-Net and to make a fair comparison with existing RGB-D SOD approaches, following these recent SOTA methods [70, 74, 96], ACFPA-Net is trained on the conventional training benchmark dataset, which consists of two parts: The training set of NJUD2K [61] dataset NJUD2K-train with 1,485 image pairs and the training set of NLPR [59] dataset NLPR-train with 700 image pairs.

Testing datasets. Six widely-used RGB-D SOD benchmark datasets are used as experimental testing datasets, which includes STEREO [58], NLPR-test [59], NJUD2K-test [61], DUTLF-D-test [66], LFSD [97] and SIP [98]. Except for DUTLF-D, other datasets are directly used for testing the performance of our proposed model and competing methods. DES [60] includes 135 images of indoor scenes captured by a Kinect camera. As the test subset of NLPR dataset, NLPR-test [59] is captured by a Kinect with a resolution of 640×480, which contains 300 natural images with multiple salient objects from 11 types of indoor and outdoor scenes under different illumination conditions. NJUD2K-test [61] is what remains of the NLPR dataset after NLPR-train has been stripped out, which contains 500 stereo image pairs with diverse objects and complex scenarios from different sources such as the internet and stereo movies, where several depth maps are estimated through an optical flow method. STEREO [58] includes 1000 stereoscopic images downloaded from the internet where the depth maps are generated from the stereo images using the SIFT flow method. SIP [98] contains 929 high-resolution RGB-D images with a high-resolution of 744×992, which covers diverse real-world scenes from various viewpoints, poses, occlusions, illuminations, and backgrounds. DUTLF-D [66] contains 1,200 paired RGB-D images captured by a Lytro camera with a resolution of 600×400. LFSD [97] is a small-scale dataset including 100 small-resolution RGB-D images and manually labeled ground truths, where the depth maps are captured via a Lytro light field camera.

4.3. Evaluation metrics

Following previous works [70, 84, 85, 96], we adopt five generally-recognized metrics for quantitative performance evaluation of the proposed method and other SOTA competitors, namely precision recall curves (PRC), mean absolute error (MAE), mean F-measure (Fm), mean E-measure (E_ξ) [105] and S-measure (S_α) [106]. Given a saliency map S and the ground truth map G, the definitions for these metrics are as follows:

The first is F-measure as a widely-used region-based similarity evaluation metric, which takes into account both precision (P) and recall (R) to assess the overall performance of the predicted results. In this work, we assess the maximum F-measure (F_β^{Max}) score across the binary predicted maps of different thresholds.

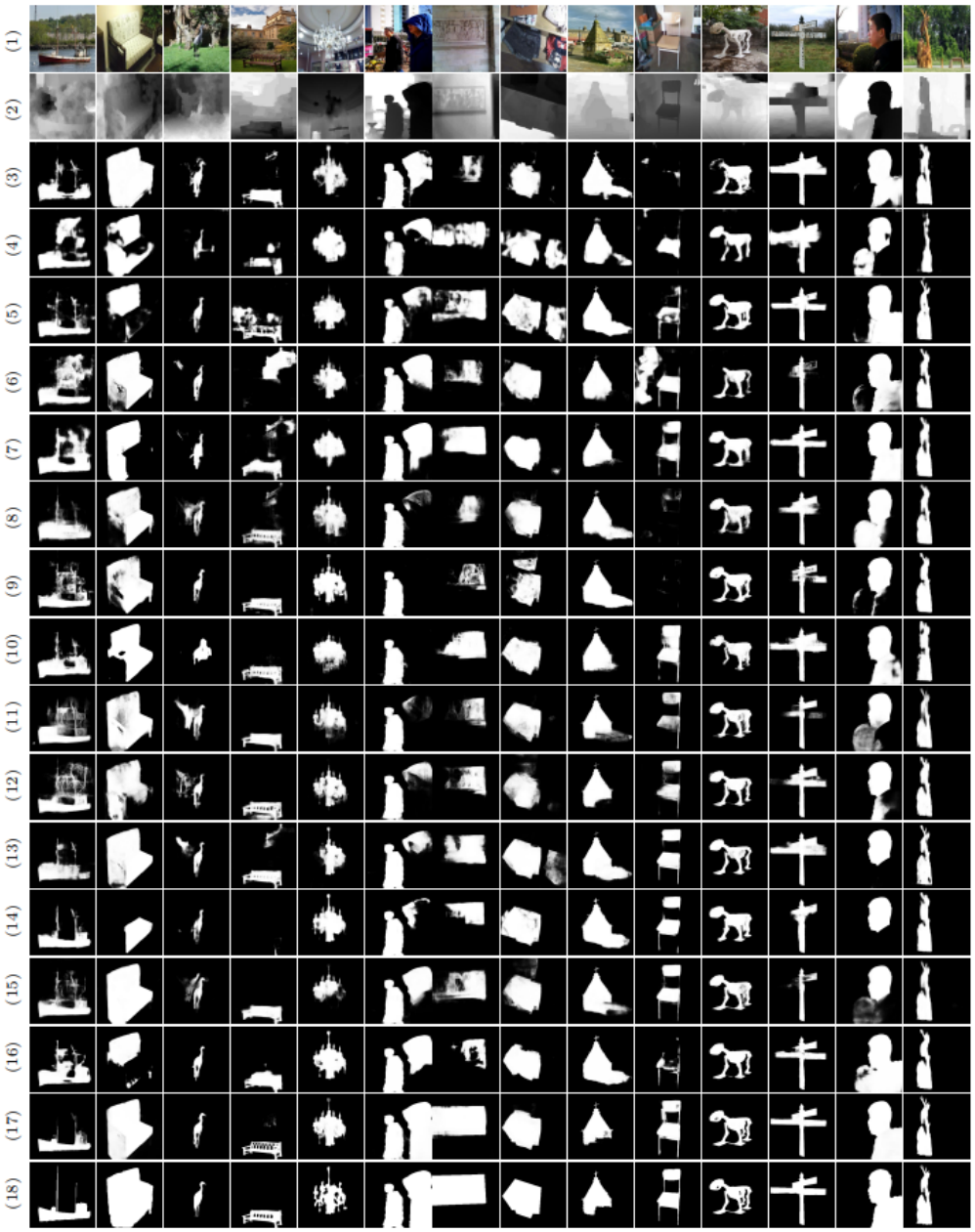

Figure 5. Visual comparison between our ACFPA-Net and SOTA RGB-D SOD methods on various challenging scenarios. (1) Image; (2) Depth; (3) CPFP [96]; (4) DMRA [66]; (5) CoNet [99]; (6) IcNet [77]; (7) JL-DCF [74]; (8) D3Net [98]; (9) RD3D [31]; (10) DFMNet [100]; (11) BBSNet [81]; (12) BTSNet [101]; (13) HAINet [69]; (14) C2DFNet [102]; (15) DIGRNet [103]; (16) CAVER [104]; (17) Our; (18) Mask.

The second is the MAE, which measures the average pixel-wise absolute difference between the

predicted saliency maps and the corresponding ground truth. It is denoted as:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S(x, y) - G(x, y)|, \tag{4.1}$$

where S is the saliency map, G is the ground truth, and W and H indicate the width and height of the saliency map respectively.

The third is S-measure (S_α), which measures the spatial structure similarities of predicted saliency maps compared to the corresponding ground truth from the perspective of region-aware and object-aware. Mathematically,

$$S_\alpha = \alpha * S_o + (1 - \alpha) * S_r, \tag{4.2}$$

where S_r denotes regional perception, S_o denotes object perception and α is set to be 0.5 to balance the region similarity S_r and object similarity S_o, as suggested in [106].

We also calculate scores of mean E-measure E_ξ^{Mean} under the same protocol as the official paper [105], which jointly captures image-level statistics and local pixel-wise correlation information to evaluate the similarity between the saliency prediction and the ground truth. Mathematically,

$$E_\xi^{Mean} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \xi_{FM}(i, j), \tag{4.3}$$

where ξ_{FM} stands for the enhanced alignment matrix at pixel location. W and H are the width and height of the saliency map.

In summary, for the four metrics above, higher F_β, E_ξ, S_α and the lower MAE score indicate better performance. Besides, we report the number of parameters and running time of each method for efficiency analysis.

Table 1. Quantitative Comparison with SOTA methods on the DES [60], NLPR [59] and LFSD [97] benchmark datasets. ↑ (↓) denotes that the higher (lower) is better. Evaluation metrics include F_β, E_ξ, S_α, and MAE score, The top 3 results are respectively marked in red, green and blue.

Method	BackBone	**Pub'Year**	**DES**				**NLPR**				**LFSD**			
			$S_\alpha\uparrow$	$F_\beta^{Max}\uparrow$	$E_\xi^{Mean}\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta^{Max}\uparrow$	$E_\xi^{Mean}\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta^{Max}\uparrow$	$E_\xi^{Mean}\uparrow$	MAE↓
CPFP [96]	VGG	CVPR-19	0.872	0.882	0.888	0.038	0.888	0.867	0.918	0.036	0.828	0.850	0.863	0.088
DMRA [66]	VGG	ICCV-19	0.897	0.885	0.933	0.030	0.899	0.879	0.940	0.031	0.846	0.856	0.893	0.076
CoNet [99]	ResNet	ECCV-20	0.910	0.915	0.945	0.028	0.907	0.848	0.933	0.031	0.862	0.848	0.901	0.071
ICNet [77]	ResNet	TIP-20	0.920	0.925	0.960	0.027	0.923	0.908	0.945	0.028	0.868	0.871	0.900	0.071
JL-DCF [74]	ResNet	TPAMI-21	0.929	0.919	0.936	0.040	0.925	0.916	0.955	0.022	0.862	0.866	0.894	0.071
D3Net [98]	VGG	TNNLS-21	0.898	0.885	0.912	0.031	0.912	0.897	0.936	0.034	0.825	0.810	0.876	0.095
RD3D [31]	ResNet	AAAI-21	0.880	0.851	0.878	0.035	0.913	0.894	0.932	0.031	0.861	0.849	0.876	0.076
DFMNet [100]	MobileNet-v2	MM-21	0.931	0.922	0.959	0.021	0.923	0.908	0.947	0.026	0.870	0.866	0.894	0.068
BBSNet [81]	ResNet50	TIP-21	0.933	0.927	0.949	0.021	0.930	0.918	0.950	0.023	0.864	0.858	0.883	0.072
BTSNet [101]	ResNet50	ICME-21	0.942	0.940	0.963	0.018	0.934	0.923	0.955	0.023	0.867	0.874	0.891	0.070
HAINet [69]	ResNet	TIP-21	0.935	0.945	0.964	0.018	0.924	0.922	0.955	0.024	0.733	0.717	0.786	0.145
C2DFNet [102]	ResNet50	TMM-22	0.921	0.937	0.945	0.020	0.927	0.926	0.958	0.021	0.863	0.926	0.897	0.065
DIGRNet [103]	ResNet50	TMM-22	0.937	0.943	0.970	0.020	0.934	0.934	0.956	0.023	0.873	0.890	0.906	0.067
CAVER [104]	ResNet50	TIP-23	0.929	0.939	0.971	0.019	0.929	0.926	0.960	0.021	0.873	0.892	0.893	0.064
MITF-Net [107]	PVT	TCVST-23	0.938	0.936	0.972	0.016	0.933	0.928	0.964	0.018	0.874	0.876	0.904	0.063
Ours	**Res50**	——	**0.940**	**0.943**	**0.978**	**0.017**	**0.930**	**0.928**	**0.961**	**0.020**	**0.899**	**0.913**	**0.930**	**0.054**

*Tables may have a footer.

Table 2. Quantitative Comparison with SOTA methods on the NJUD2K [61], SIP [98] and STEREO [58] benchmark datasets. ↑ (↓) denotes that the higher (lower) is better. Evaluation metrics include F_β, E_ξ, S_α, and MAE score, The top 3 results are respectively marked in red, green and blue.

Method	BackBone	Pub'Year	NJUD2K				SIP				STEREO			
			$S_\alpha\uparrow$	$F_\beta^{Max}\uparrow$	$E_\xi^{Mean}\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta^{Max}\uparrow$	$E_\xi^{Mean}\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta^{Max}\uparrow$	$E_\xi^{Mean}\uparrow$	MAE↓
CPFP [96]	VGG	CVPR-19	0.879	0.877	0.910	0.053	0.850	0.851	0.893	0.064	0.879	0.874	0.911	0.051
DMRA [66]	VGG	ICCV-19	0.886	0.886	0.920	0.051	0.806	0.821	0.844	0.085	0.835	0.895	0.879	0.066
CoNet [99]	ResNet	ECCV-20	0.894	0.872	0.924	0.047	0.858	0.842	0.909	0.063	0.905	0.909	0.941	0.037
ICNet [77]	ResNet	TIP-20	0.894	0.891	0.913	0.052	0.854	0.857	0.900	0.069	0.903	0.898	0.926	0.045
JL-DCF [74]	ResNet	TPAMI-21	0.903	0.903	0.935	0.043	0.879	0.885	0.918	0.051	0.905	0.901	0.936	0.042
D3Net [98]	VGG	TNNLS-21	0.900	0.900	0.876	0.041	0.860	0.861	0.897	0.063	0.891	0.891	0.921	0.046
RD3D [31]	ResNet	AAAI-21	0.928	0.928	0.943	0.033	0.843	0.840	0.870	0.069	0.899	0.888	0.916	0.046
DFMNet [100]	MobileNet-v2	MM-21	0.895	0.889	0.925	0.045	0.883	0.887	0.914	0.051	0.898	0.892	0.928	0.045
BBSNet [81]	ResNet50	TIP-21	0.921	0.920	0.938	0.035	0.879	0.883	0.906	0.055	0.908	0.903	0.928	0.041
BTSNet [101]	ResNet50	ICME-21	0.910	0.901	0.927	0.037	0.896	0.901	0.924	0.044	0.914	0.911	0.938	0.038
HAINet [69]	ResNet	TIP-21	0.841	0.831	0.882	0.069	0.880	0.892	0.916	0.053	0.907	0.906	0.936	0.040
C2DFNet [102]	ResNet50	TMM-22	0.907	0.918	0.936	0.038	0.871	0.895	0.912	0.052	0.902	0.911	0.936	0.038
DIGRNet [103]	ResNet50	TMM-22	0.932	0.943	0.954	0.028	0.885	0.912	0.918	0.053	0.916	0.921	0.943	0.038
CAVER [104]	ResNet50	TIP-23	0.921	0.932	0.953	0.032	0.893	0.913	0.928	0.043	0.913	0.920	0.949	0.033
MITF-Net [107]	PVT	TCVST-23	0.923	0.926	0.953	0.030	0.899	0.913	0.936	0.040	0.909	0.910	0.946	0.034
Ours	**Res50**	——	**0.933**	**0.940**	**0.952**	**0.028**	**0.905**	**0.927**	**0.943**	**0.038**	**0.924**	**0.936**	**0.954**	**0.030**

*Tables may have a footer.

4.4. Comparison with SOTA methods

To demonstrate the effectiveness of our proposed ACFPA-Net model, we qualitatively and quantitatively compare it with 15 SOTA RGB-D SOD methods on six benchmark datasets mentioned above, including MITFNet [107], CAVER [104], DIGRNet [103], C2DFNet [102], HAINet [69], BTSNet [101], BBSNet [81], DFMNet [100], RD3D [31], D3Net [98], JL-DCF [74], ICNet [77], CoNet [99], DMRA [66] and CPFP [96]. All competitors are recent deep learning-based RGB-D SOD methods. For fair comparisons, all results for these 15 compared methods are either reproduced by the authorized codes under the default settings or are directly provided by authors.

Qualitative evaluation and discussion For visual comparisons between our approach and the baseline methods, as shown in Figure 5, for several visualization results in various challenging scenes including multiple objects (column six), small objects (columns three, four and five), big objects (columns one, two, five and twelve), complex background (columns four, eight, ten and fourteen), low contrast (columns four, five, seven and thirteen), Occlusive object (column nine), complex objects (columns ten and eleven), it can be clearly seen that the proposed ACFPA-Net can consistently produce more accurate and complete saliency maps with sharp boundaries and coherent details, which achieves better detection performance of the salient objects. These samples given in Figure 5 belong to the conventional benchmark datasets, including DES [60], NLPR [59], NJUD2K [61], STEREO [58] and SIP [98].

Quantitative comparison and analysis The detailed quantitative performances of the proposed ACFPA-Net and previous competitive methods on DES, NLPR and LFSD datasets can be found in Table 1, and Table 2 shows the quantitative comparison in terms of four evaluation metrics on NJUD2K, SIP, and STEREO datasets. These evaluation values in both Table 1 and Table 2 show that our ACFPA-Net can more accurately and completely detect salient objects in complex scenes by

using global semantic relations and multi-scale detail information. Compared with the other 15 methods, our carefully designed ACFPA-Net ranks in the top three in terms of S-measure, E-measure and F-measure. For example, on the SIP dataset, compared to the second best method (MITF-Net), the percentage gain of E-measure reaches 0.7%, the percentage gain of MAE score reaches 0.2% and the percentage gain of S-measure reaches 1.5%. On the STEREO dataset, the minimum percentage gain measured by S-measure reaches 0.8%, while the minimum percentage gain measured by F-measure reaches 1.5%.

Model complexity As shown in Table 3, we compare our model complexity with some recent competitive ones, and it can be seen that our model has a relative disadvantage in the aspect of expenses; however, this is not the focus of this article, and Tables 1 and 2 have proved our value in performance. Of course, model lightweighting is also important research with wide application value, and we will work on more efficient methods in the future.

Table 3. Complexity comparison with SOTA methods.

Methods	TriTransNet [108]	SPNet [88]	CIRNet [109]	SwinNet [110]	CAVER [104]	PICRNet [111]	Ours
Pub'year	ACMMM-21	ICCV-21	TIP-22	TCSVT-22	TIP-23	ACMMM-23	—
parameter(Mb)	1290	670	394	705	214	426	625

4.5. Ablation study

We have provided comprehensive ablation studies to validate the effectiveness of each key component (PDC and CPFI) employed in our ACFPA-Net. The quantitative results for each module combination are shown in Table 4. For a fair comparison, the benchmark of this paper is to directly add RGB features and depth features as fusion features and input them into the CMFD module. We gradually add different components to the baseline.

Table 4. Ablation study of the proposed modules. CPFI and PDC denote the cross-modal pyramid feature interaction module and the pyramid dilated convolution module,respectively.

Method	NJUD				SIP				STERE			
	$S_\alpha\uparrow$	$F_\beta^{Max}\uparrow$	$E_\xi^{Mean}\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta^{Max}\uparrow$	$E_\xi^{Mean}\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta^{Max}\uparrow$	$E_\xi^{Mean}\uparrow$	MAE↓
baseline	0.652	0.597	0.651	0.242	0.486	0.425	0.663	0.325	0.549	0.578	0.612	0.238
+PDC	0.713	0.720	0.776	0.159	0.592	0.591	0.711	0.234	0.662	0.631	0.736	0.199
+CPFI	0.866	0.892	0.884	0.063	0.741	0.740	0.814	0.135	0.827	0.841	0.851	0.085
+CBAM+PDC	0.915	0.916	0.940	0.035	0.891	0.908	0.932	0.046	0.904	0.918	0.940	0.039
+RCBAM+PDC	0.921	0.929	0.947	0.033	0.898	0.911	0.937	0.045	0.909	0.922	0.945	0.036
+CPFI+PDC	**0.933**	**0.940**	**0.952**	**0.028**	**0.905**	**0.927**	**0.943**	**0.038**	**0.924**	**0.936**	**0.954**	**0.030**
+CIM+PDC	0.924	0.930	0.938	0.030	0.899	0.916	0.938	0.041	0.911	0.927	0.942	0.033
+CMIB+PDC	0.926	0.933	0.947	0.029	**0.905**	0.920	0.940	**0.038**	0.919	0.930	0.948	0.029

The Significance of the PDC module. To prove the effectiveness of the PDC, quantitative comparisons are illustrated in Table 4. It can be observed that after adding the PDC module to the first line, there is a significant improvement in various evaluation indicators. This also confirms our idea that by incorporating the PDC module after high-level features, we can capture broader and deeper

semantic features, providing global semantic support for the decoding stage. This clarity demonstrates the effectiveness of the CPFI module.

The effectiveness of CPFI module. First, to study the importance of the CPFI module, we listed the experiments on the NJUD, SIP and STEREO datasets, as shown in Table 3. Compared to the benchmark (first row), after adding the CPFI module to replace the direct addition operation (fifth row), the indicators have improved on multiple datasets such as NJUD and STEREO. This is reasonable because CPFI can enhance the refinement effect of single modal features (RGB features, depth features) through the multiscale convolution and RCBAM and multilevel interaction between RGB features and depth features. The last row of Table 4 adds both CPFI and PDC modules on top of the benchmark. It can be observed that each method improves performance and, when combined, we achieve the best results. Second, we explore the effectiveness of the CPFI structure.The key structures of CPFI are RCBAM and multi-scale aggregation. We replace CPFI with single-scale CBAM (third row) and single-scale RCBAM (fourth row) with the decoder PDC. The comparison between the third and fourth rows can demonstrate that adding residual connections to CBAM is positive for the model. The comparison between the fifth and fourth rows can demonstrate the effectiveness of multi-scale aggregation. In addition, we also compare CPFI with other aggregation modules, and CPFI gets promising results, as shown in the fifth to seventh rows. CIM and CMIB are two SOTA methods SPNet [88] and DIGRNet [103] aggregation modules, respectively.

4.6. Failure cases

The proposed model has good detection performance in most cases. However, as shown in Figure 6, there are some failure cases. The first row can't effectively avoid the interference of occlusion. The second row fails to highlight the object due to the fact that the object is at the edge of the frame and is incomplete. The third column has incomplete detection results due to the complexity of the scene and the difficulty in distinguishing the object from other pedestrians. Similar failed detections are also similarly seen in other SOTA RGB-D SOD methods, as shown in the fifth and sixth columns. In addition to the above reasons, we believe that low-quality depth maps (third row) are also important in affecting the performance of the model. These extremely challenging issues will also be the focus of our future research.

5. Conclusions

In this paper, we propose an ACFPA-Net for RGB-D SOD. To effectively balance high accuracy and efficiency, we designed a CPFI module to integrate multilevel features from both RGB and depth modalities. Moreover, for improving the intra- and inter-modality aggregation compatibility when fusing the information of the two RGB and depth modalities, we introduced a CMFD to aggregate these fused features in an encoder network to generate the final prediction results. Experiments on six benchmark datasets demonstrated that the proposed ACFPA-Net achieves competitive performance over 15 SOTA RGB-D SOD models, under four widely used evaluation metrics.

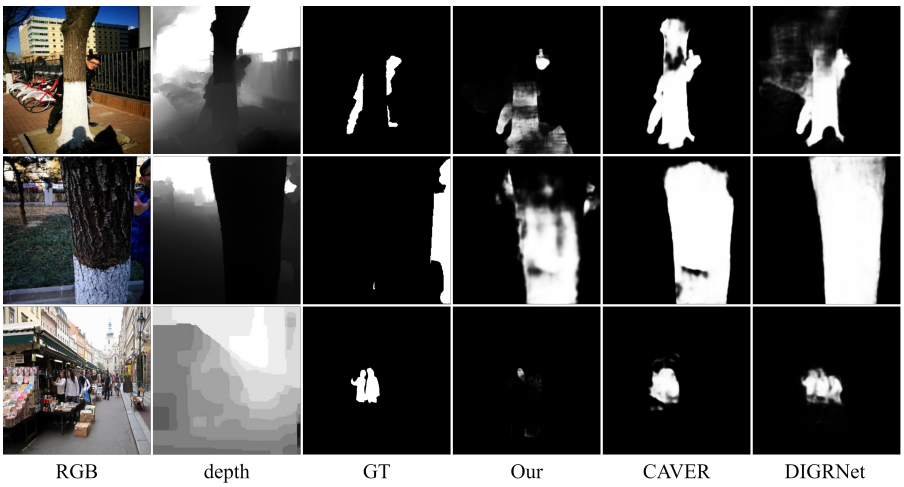

Figure 6. Failure cases.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. Y. Zhao, Y. Peng, Saliency-guided video classification via adaptively weighted learning, in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, (2017), 847–852. https://doi.org/10.1109/ICME.2017.8019343
2. X. Hu, Y. Wang, J. Shan, Automatic recognition of cloud images by using visual saliency features, *IEEE Geosci. Remote Sens. Lett.*, **12** (2015), 1760–1764. https://doi.org/10.1109/LGRS.2015.2424531
3. J. C. Ni, Y. Luo, D. Wang, J. Liang, Q. Zhang, Saliency-based sar target detection via convolutional sparse feature enhancement and bayesian inference, *IEEE Trans. Geosci. Remote Sens.*, **61** (2023), 1–15. https://doi.org/10.1109/TGRS.2023.3237632
4. Z. Yu, Y. Zhuge, H. Lu, L. Zhang, Joint learning of saliency detection and weakly supervised semantic segmentation, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 7222–7232. https://doi.org/10.1109/ICCV.2019.00732
5. S. Lee, M. Lee, J. Lee, H. Shim, Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 5491–5501. https://doi.org/10.1109/CVPR46437.2021.00545

6. W. Feng, R. Han, Q. Guo, J. Zhu, S, Wang, Dynamic saliency-aware regularization for correlation filter-based object tracking, *IEEE Trans.n Image Process.*, **28** (2019), 3232–3245. https://doi.org/10.1109/TIP.2019.2895411
7. J. Y. Zhu, J. Wu, Y. Xu, E. Chang, Z. Tu, Unsupervised object class discovery via saliency-guided multiple class learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** (2015), 862–875. https://doi.org/10.1109/TPAMI.2014.2353617
8. S. Wei, L. Liao, J. Li, Q. Zheng, F. Yang, Y. Zhao, Saliency inside: Learning attentive cnns for content-based image retrieval, *IEEE Trans. Image Process.*, **28** (2019), 4580–4593. https://doi.org/10.1109/TIP.2019.2913513
9. A. Kim, R. M. Eustice, Real-time visual slam for autonomous underwater hull inspection using visual saliency, *IEEE Trans. Rob.*, **29** (2013), 719–733. https://doi.org/10.1109/TRO.2012.2235699
10. R. Li, C. H. Wu, S. Liu, J. Wang, G. Wang, G. Liu, B. Zeng, SDP-GAN: Saliency detail preservation generative adversarial networks for high perceptual quality style transfer, *IEEE Trans. Image Process.*, **30** (2021), 374–385. https://doi.org/10.1109/TIP.2020.3036754
11. L. Jiang, M. Xu, X. Wang, L. Sigal, Saliency-guided image translation, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 16504–16513. https://doi.org/10.1109/CVPR46437.2021.01624
12. S. Li, M. Xu, Y. Ren, Z. Wang, Closed-form optimization on saliency-guided image compression for HEVC-MSP, *IEEE Trans. Multimedia*, **20** (2018), 155–170. https://doi.org/10.1109/TMM.2017.2721544
13. Y. Patel, S. Appalaraju, R. Manmatha, Saliency driven perceptual image compression, in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (2021), 227–231. https://doi.org/10.1109/WACV48630.2021.00027
14. C. Yang, L. Zhang, H. Lu, X. Ruan, M. H. Yang, Saliency detection via graph-based manifold ranking, in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, (2013), 3166–3173. https://doi.org/10.1109/CVPR.2013.407
15. W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 2814–2821. https://doi.org/10.1109/CVPR.2014.360
16. K. Shi, K. Wang, J. Lu, L. Lin, Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors, in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, (2013), 2115–2122. https://doi.org/10.1109/CVPR.2013.275
17. M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, S. M. Hu., Global contrast based salient region detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** (2015), 569–582. https://doi.org/10.1109/CVPR.2011.5995344 https://doi.org/10.1109/TPAMI.2014.2345401
18. W. C. Tu, S. He, Q. Yang, S. Y. Chien, Real-time salient object detection with a minimum spanning tree, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2334–2342. https://doi.org/10.1109/CVPR.2016.256

19. R. Zhao, W. Ouyang, H. Li, X, Wang, Saliency detection by multi-context deep learning, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 1265–1274. https://doi.org/10.1109/CVPR.2015.7298731
20. G. Li, Y. Yu, Visual saliency based on multiscale deep features, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 5455–5463.
21. L. Wang, H. Lu, X. Ruan, M. H. Yang, Deep networks for saliency detection via local estimation and global search, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 3183–3192. https://doi.org/10.1109/CVPR.2015.7298938
22. Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, P. M. Jodoin, Non-local deep features for salient object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 6593–6601. https://doi.org/10.1109/CVPR.2017.698
23. P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: Aggregating multi-level convolutional features for salient object detection, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 202–211. https://doi.org/10.1109/ICCV.2017.31
24. Q. Hou, M. M. Cheng, X. Hu, A. Borji, Z. Tu, P. Torr, Deeply supervised salient object detection with short connections, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 5300–5309. https://doi.org/10.1109/CVPR.2017.563
25. W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, R. Yang, Salient object detection in the deep learning era: An in-depth survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 3239–3259. https://doi.org/10.1109/TPAMI.2021.3051099
26. A. Borji, M. M. Cheng, Q. Hou, H. Jiang, J. Li, Salient object detection: A survey, *Comput. Vis. Media*, **5** (2019), 117–150. https://doi.org/10.1007/s41095-019-0149-9
27. T. Zhou, D. P. Fan, M. M. Cheng, J. Shen, L. Shao, RGB-D salient object detection: A survey, *Comput. Vis. Media*, **7** (2021), 37–69. https://doi.org/10.1007/s41095-020-0199-z
28. X. Song, D. Zhou, W. Li, Y. Dai, L. Liu, H. Li, et al., WAFP-Net: Weighted attention fusion based progressive residual learning for depth map super-resolution, *IEEE Trans. Multimedia*, **24** (2022), 4113–4127. https://doi.org/10.1109/TMM.2021.3118282
29. P. F. Proença, Y. Gao, Splode: Semi-probabilistic point and line odometry with depth estimation from RGB-D camera motion, in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (2017), 1594–1601. https://doi.org/10.1109/IROS.2017.8205967
30. X. Xing, Y. Cai, T. Lu, Y. Yang, D. Wen, Joint self-supervised monocular depth estimation and SLAM, in *2022 26th International Conference on Pattern Recognition (ICPR)*, (2022), 4030–4036. https://doi.org/10.1109/ICPR56361.2022.9956576
31. Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, H. Du, RGB-D salient object detection via 3d convolutional neural networks, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2021), 1063–1071. https://doi.org/10.1609/aaai.v35i2.16191
32. F. Wang, J. Pan, S. Xu, J. Tang, Learning discriminative cross-modality features for RGB-D saliency detection, *IEEE Trans. Image Process.*, **31** (2022), 1285–1297. https://doi.org/10.1109/TIP.2022.3140606

33. Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, C. Demonceaux, Hidanet: RGB-D salient object detection via hierarchical depth awareness., *IEEE Trans. Image Process.*, **32** (2023), 2160–2173. https://doi.org/10.1109/TIP.2023.3263111

34. J. Zhang, Q. Liang, Q. Guo, J. Yang, Q. Zhang, Y. Shi, R2net: Residual refinement network for salient object detection, *Image Vision Comput.*, **120** (2022), 104423. https://doi.org/10.1016/j.imavis.2022.104423

35. R. Shigematsu, D. Feng, S. You, N. Barnes, Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features, in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, (2017), 2749–2757.

36. L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, **20** (1998), 1254–1259. https://doi.org/10.1109/34.730558

37. C. Yang, L. Zhang, H. Lu, Graph-regularized saliency detection with convex-hull-based center prior, *IEEE Signal Process. Lett.*, **20** (2013), 637–640. https://doi.org/10.1109/LSP.2013.2260737

38. P. Jiang, H. Ling, J. Yu, J. Peng, Salient region detection by ufo: Uniqueness, focusness and objectness, in *2013 IEEE International Conference on Computer Vision*, (2013), 1976–1983.

39. R. S. Srivatsa, R. V. Babu, Salient object detection via objectness measure, in *2015 IEEE International Conference on Image Processing (ICIP)*, (2015), 4481–4485. https://doi.org/10.1109/ICIP.2015.7351654

40. C. Scharfenberger, A. Wong, K. Fergani, J. S. Zelek, D. A. Clausi, Statistical textural distinctiveness for salient region detection in natural images, in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, (2013), 979–986. https://doi.org/10.1109/CVPR.2013.131

41. A. Borji, M. M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, *IEEE Trans. Image Process.*, **24** (2015), 5706–5722. https://doi.org/10.1109/TIP.2015.2487833

42. J. Han, D. Zhang, G. Cheng, N. Liu, D. Xu, Advanced deep-learning techniques for salient and category-specific object detection: A survey, *IEEE Signal Process. Mag.*, **35** (2018), 84–100. https://doi.org/10.1109/MSP.2017.2749125

43. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Adv. Neural Inf. Process. Syst.*, **2012** (2012), 25. https://doi.org/10.1145/3065386

44. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

45. N. Liu, J. Han, M. H. Yang, Picanet: Learning pixel-wise contextual attention for saliency detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 3089–3098. https://doi.org/10.1109/CVPR.2018.00326

46. S. Chen, X. Tan, B. Wang, X. Hu, Reverse attention for salient object detection, in *Proceedings of the European conference on computer vision (ECCV)*, (2018), 234–250. https://doi.org/10.1007/978-3-030-01240-3_15

47. J. J. Liu, Q. Hou, M. M. Cheng, J. Feng, J. Jiang, A simple pooling-based design for real-time salient object detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 3912–3921. https://doi.org/10.1109/CVPR.2019.00404

48. Y. Pang, X. Zhao, L. Zhang, H. Lu, Multi-scale interactive network for salient object detection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 9410–9419. https://doi.org/10.1109/CVPR42600.2020.00943

49. Q. Hou, M. M. Cheng, X. Hu, A. Borji, Z. Tu, P. H. S. Torr, Deeply supervised salient object detection with short connections, *IEEE Trans. Pattern Anal. Mach. Intell.*, **41** (2019), 815–828. https://doi.org/10.1109/CVPR.2017.563 https://doi.org/10.1109/TPAMI.2018.2815688

50. X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, Basnet: Boundary-aware salient object detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 7471–7481. https://doi.org/10.1109/CVPR.2019.00766

51. P. Zhang, W. Liu, H. Lu, C. Shen, Salient object detection with lossless feature reflection and weighted structural loss, *IEEE Trans. Image Process.*, **28** (2019), 3048–3060. https://doi.org/10.1109/TIP.2019.2893535

52. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in *3rd International Conference on Learning Representations*, 2015.

53. W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 1448–1457. https://doi.org/10.1109/CVPR.2019.00154

54. T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 3080–3089. https://doi.org/10.1109/CVPR.2019.00320

55. S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, Y. Fu, Reverse attention-based residual network for salient object detection, *IEEE Trans. Image Process.*, **29** (2020), 3763–3776. https://doi.org/10.1109/TIP.2020.2965989

56. M. Feng, H. Lu, E. Ding, Attentive feedback network for boundary-aware salient object detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 1623–1632. https://doi.org/10.1109/CVPR.2019.00172

57. J. Zhao, J. J. Liu, D. P. Fan, Y. Cao, J. Yang, M. M. Cheng, Egnet: Edge guidance network for salient object detection, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 8778–8787. https://doi.org/10.1109/ICCV.2019.00887

58. Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (2012), 454–461.

59. H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: A benchmark and algorithms, in *Computer Vision–ECCV 2014: 13th European Conference*, (2014), 92–109. https://doi.org/10.1007/978-3-319-10578-9_7

60. Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, in *Proceedings of international conference on internet multimedia computing and service*, (2014), 23–27. https://doi.org/10.1145/2632856.2632866

61. R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in *2014 IEEE International Conference on Image Processing (ICIP)*, (2014), 1115–1119. https://doi.org/10.1109/ICIP.2014.7025222

62. J. Ren, X. Gong, L. Yu, W. Zhou, M. Y. Yang, Exploiting global priors for rgb-d saliency detection, in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2015), 25–32. https://doi.org/10.1109/CVPRW.2015.7301391

63. A. Wang, M. Wang, RGB-D salient object detection via minimum barrier distance transform and saliency fusion, *IEEE Signal Process. Lett.*, **24** (2017), 663–667. https://doi.org/10.1109/LSP.2017.2688136

64. R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, S. Kwong, Going from RGB to RGBD saliency: A depth-guided transformation model, *IEEE Trans. Cyber.*, **50** (2020), 3627–3639. https://doi.org/10.1109/TCYB.2019.2932005

65. L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, Q. Yang, RGBD salient object detection via deep fusion, *IEEE Trans. Image Process.*, **26** (2017), 2274–2285. https://doi.org/10.1109/TIP.2017.2682981

66. Y. Piao, W. Ji, J. Li, M. Zhang, H. Lu, Depth-induced multi-scale recurrent attention network for saliency detection, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 7253–7262. https://doi.org/10.1109/ICCV.2019.00735

67. N. Liu, N. Zhang, J. Han, Learning selective self-mutual attention for RGB-D saliency detection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 13753–13762. https://doi.org/10.1109/CVPR42600.2020.01377

68. C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, et al., ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection, *IEEE Trans. Cyber.*, **51** (2021), 88–100. https://doi.org/10.1109/TCYB.2020.2969255

69. G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, H. Ling, Hierarchical alternate interaction network for RGB-D salient object detection, *IEEE Trans. Image Process.*, **30** (2021), 3528–3542. https://doi.org/10.1109/TIP.2021.3062689

70. Y. H. Wu, Y. Liu, J. Xu, J. W. Bian, Y. C. Gu, M. M. Cheng, MobileSal: Extremely efficient RGB-D salient object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 10261–10269. https://doi.org/10.1109/TPAMI.2021.3134684

71. N. Huang, Y. Yang, D. Zhang, Q. Zhang, J. Han, Employing bilinear fusion and saliency prior information for RGB-D salient object detection, *IEEE Trans. Multimedia*, **24** (2022), 1651–1664. https://doi.org/10.1109/TMM.2021.3069297

72. X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, H. Qin, Data-level recombination and lightweight fusion scheme for RGB-D salient object detection, *IEEE Trans. Image Process.*, **30** (2021), 458–471. https://doi.org/10.1109/TIP.2020.3037470

73. X. Zhao, L. Zhang, Y. Pang, H. Lu, L. Zhang, A single stream network for robust and real-time RGB-D salient object detection, in *Computer Vision—ECCV 2020: 16th European Conference*, (2020), 646–662. https://doi.org/10.1007/978-3-030-58542-6_39

74. K. Fu, D. P. Fan, G. P. Ji, Q. Zhao, JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 3049–3059. https://doi.org/10.1109/CVPR42600.2020.00312

75. J. Han, H. Chen, N. Liu, C. Yan, X. Li, CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion, *IEEE Trans. Cyber.*, **48** (2018), 3171–3183. https://doi.org/10.1109/TCYB.2017.2761775

76. N. Wang, X. Gong, Adaptive fusion for RGB-D salient object detection, *IEEE Access*, **7** (2019), 55277–55284. https://doi.org/10.1109/ACCESS.2019.2913107

77. G. Li, Z. Liu, H. Ling, ICNet: Information conversion network for RGB-D based salient object detection, *IEEE Trans. Image Process.*, **29** (2020), 4873–4884. https://doi.org/10.1109/TIP.2020.2976689

78. H. Chen, Y. Li, Progressively complementarity-aware fusion network for RGB-D salient object detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 3051–3060. https://doi.org/10.1109/CVPR.2018.00322

79. M. Zhang, W. Ren, Y. Piao, Z. Rong, H.Lu, Select, supplement and focus for RGB-D saliency detection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 3469–3478. https://doi.org/10.1109/CVPR42600.2020.00353

80. C. Chen, J. Wei, C. Peng, H. Qin, Depth-quality-aware salient object detection, *IEEE Trans. Image Process.*, **30** (2021), 2350–2363. https://doi.org/10.1109/TIP.2021.3052069

81. Y. Zhai, D. P. Fan, J. Yang, A. Borji, L. Shao, J. Han, L. Wang, Bifurcated backbone strategy for RGB-D salient object detection, *IEEE Trans. Image Process.*, **30** (2021), 8727–8742. https://doi.org/10.1109/TIP.2021.3116793

82. W. D. Jin, J. Xu, Q. Han, Y. Zhang, M. M. Cheng, CDNet: Complementary depth network for RGB-D salient object detection, *IEEE Trans. Image Process.*, **30** (2021), 3376–3390. https://doi.org/10.1109/TIP.2021.3060167

83. Z. Zhang, Z. Lin, J. Xu, W. D. Jin, S. P. Lu, D. P. Fan, Bilateral attention network for RGB-D salient object detection, *IEEE Trans. Image Process.*, **30** (2021), 1949–1961. https://doi.org/10.1109/TIP.2021.3049959

84. H. Chen, Y. Li, Three-stream attention-aware network for RGB-D salient object detection, *IEEE Trans. Image Process.*, **28** (2019), 2825–2835. https://doi.org/10.1109/TIP.2019.2891104

85. J. Zhang, D. P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, et al., Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders, iIn *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 8579–8588. https://doi.org/10.1109/CVPR42600.2020.00861

86. A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, S. Lyu, Cascade graph neural networks for RGB-D salient object detection, in *Computer Vision—ECCV 2020: 16th European Conference*, (2020), 346–364. https://doi.org/10.1007/978-3-030-58610-2_21

87. B. Jiang, Z. Zhou, X. Wang, J. Tang, B. Luo, CmSalGAN: RGB-D salient object detection with cross-view generative adversarial networks, *IEEE Trans. Multimedia*, **23** (2021), 1343–1353. https://doi.org/10.1109/TMM.2020.2997184

88. T. Zhou, H. Fu, G. Chen, Y. Zhou, D. P. Fan, L. Shao, Specificity-preserving RGB-D saliency detection, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 4661–4671. https://doi.org/10.1109/ICCV48922.2021.00464

89. T. Zhou, Y. Zhou, C. Gong, J. Yang, Y. Zhang, Feature aggregation and propagation network for camouflaged object detection, *IEEE Trans. Image Process.*, **31** (2022), 7036–7047. https://doi.org/10.1109/TIP.2022.3217695

90. M. Song, W. Song, G. Yang, C. Chen, Improving RGB-D salient object detection via modality-aware decoder, *IEEE Trans. Image Process.*, **31** (2022), 6124–6138. https://doi.org/10.1109/TIP.2022.3205747

91. Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, et al., Ce-net: Context encoder network for 2d medical image segmentation, *IEEE Trans. Med. Imaging*, **38** (2019), 2281–2292. https://doi.org/10.1109/TMI.2019.2903562

92. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

93. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 2091–2106. https://doi.org/10.1109/TCSVT.2021.3082939

94. K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** (2015), 1904–1916. https://doi.org/10.1007/978-3-319-10578-9_23

95. I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in *7th International Conference on Learning Representations*, 2019.

96. J. X. Zhao, Y. Cao, D. P. Fan, M. M. Cheng, X. Y. Li, L. Zhang, Contrast prior and fluid pyramid integration for RGB-D salient object detection, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 3922–3931.

97. N. Li, J. Ye, Y. Ji, H. Ling, J. Yu, Saliency detection on light field, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 2806–2813. https://doi.org/10.1109/CVPR.2014.359

98. D. P. Fan, Z. Lin, Z. Zhang, M. Zhu, M. M. Cheng, Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks, *IEEE Trans. Neural Networks Learn. Syst.*, **32** (2021), 2075–2089. https://doi.org/10.1109/TNNLS.2020.2996406

99. W. Ji, J. Li, M. Zhang, Y. Piao, H. Lu, Accurate RGB-D salient object detection via collaborative learning, in *Computer Vision—ECCV 2020: 16th European Conference*, (2020), 52–69. https://doi.org/10.1007/978-3-030-58523-5_4

100. W. Zhang, G. P. Ji, Z. Wang, K. Fu, 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. https://doi.org/10.1145/3474085.3475240

101. W. Zhang, Y. Jiang, K. Fu, Q. Zhao, BTS-Net: Bi-directional transfer-and-selection network for RGB-D salient object detection, in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, (2021), 1–6. https://doi.org/10.1109/ICME51207.2021.9428263

102. M. Zhang, S. Yao, B. Hu, Y. Piao, W. Ji, C^2DFNet: Criss-cross dynamic filter network for rgb-d salient object detection, *IEEE Trans. Multimedia*, **2022** (2022), 1–13.

103. X. Cheng, X. Zheng, J. Pei, H. Tang, Z. Lyu, C. Chen, Depth-induced gap-reducing network for RGB-D salient object detection: An interaction, guidance and refinement approach, *IEEE Trans. Multimedia*, **2022** (2022).

104. Y. Pang, X. Zhao, L. Zhang, H. Lu, Caver: Cross-modal view-mixed transformer for bi-modal salient object detection, *IEEE Trans. Image Process.*, **32** (2023), 892–904. https://doi.org/10.1109/TIP.2023.3234702

105. D. P. Fan, C. Gong, Y. Cao, B. Ren, M. M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, (2018), 698–704. https://doi.org/10.24963/ijcai.2018/97

106. D. P. Fan, M. M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 4558–4567. https://doi.org/10.1109/ICCV.2017.487

107. G. Chen, F. Shao, X. Chai, H. Chen, Q. Jiang, X. Meng, Y. S. Ho, Modality-induced transfer-fusion network for RGB-D and RGB-T salient object detection, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 1787–1801. https://doi.org/10.1109/TCSVT.2022.3215979

108. Z. Liu, Y. Wang, Z. Tu, Y. Xiao, B. Tang, Tritransnet, in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. https://doi.org/10.1145/3474085.3475601

109. R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, Y. Zhao, CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection, *IEEE Trans. Image Process.*, **31** (2022), 6800–6815. https://doi.org/10.1109/TIP.2022.3216198

110. Z. Liu, Y. Tan, Q. He, Y. Xiao, Swinnet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 4486–4497. https://doi.org/10.1109/TCSVT.2021.3127149

111. R. Cong, H. Liu, C. Zhang, W. Zhang, F. Zheng, R. Song, 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. https://doi.org/10.1145/3581783.3611982

AIMS Press