Bilateral Attention Network for RGB-D Salient Object Detection

Zhao Zhang, Zheng Lin, Jun Xu, *Member, IEEE*, Wen-Da Jin, Shao-Ping Lu, *Member, IEEE*, and Deng-Ping Fan, *Member, IEEE*

***Abstract*—RGB-D salient object detection (SOD) aims to segment the most attractive objects in a pair of cross-modal RGB and depth images. Currently, most existing RGB-D SOD methods focus on the foreground region when utilizing the depth images. However, the background also provides important information in traditional SOD methods for promising performance. To better explore salient information in both foreground and background regions, this paper proposes a Bilateral Attention Network (BiANet) for the RGB-D SOD task. Specifically, we introduce a Bilateral Attention Module (BAM) with a complementary attention mechanism: foreground-first (FF) attention and background-first (BF) attention. The FF attention focuses on the foreground region with a gradual refinement style, while the BF one recovers potentially useful salient information in the background region. Benefited from the proposed BAM module, our BiANet can capture more meaningful foreground and background cues, and shift more attention to refining the uncertain details between foreground and background regions. Additionally, we extend our BAM by leveraging the multi-scale techniques for better SOD performance. Extensive experiments on six benchmark datasets demonstrate that our BiANet outperforms other state-of-the-art RGB-D SOD methods in terms of objective metrics and subjective visual comparison. Our BiANet can run up to 80 *fps* on 224 × 224 RGB-D images, with an NVIDIA GeForce RTX 2080Ti GPU. Comprehensive ablation studies also validate our contributions.**

***Index Terms*—Bilateral attention, salient object detection, RGB-D image.**

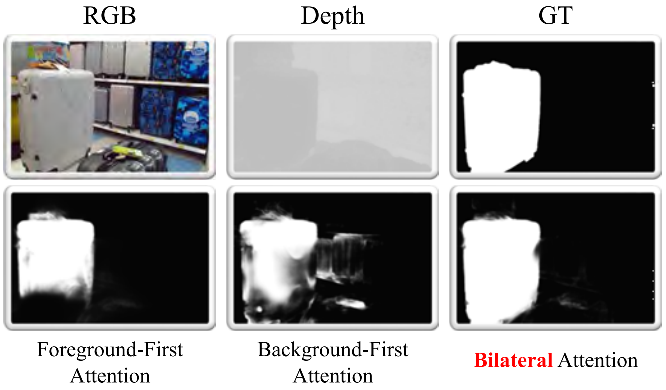

Fig. 1. **Comparison of RGB-D SOD results by Foreground-First, Background-First, and our Bilateral attention mechanisms.** Depth information provides rich foreground and background relationships. Paying more attention to foreground helps to predict high-confidence foreground objects, but may produce incomplete results. Focusing more on background finds more complete objects, but may introduce unexpected noise. Our BiANet jointly explores foreground and background cues, and achieves complete foreground prediction with little background noise.

I. Introduction

FOR understanding complex scenes in real time, humans are able to filter visually distinctive, so-called salient, subset of the available visual information before further processing [32], [69]. This capability has long been studied by researchers in physiology, cognitive psychology, computer vision, *etc.* [10], [31], [89]. A salient object can be distinctive from its neighbors in color, shape, distance, *etc.* [4], [47]. Capturing the attention-grabbing objects first has been proved to be effective in wide vision applications, such as visual tracking [41], [49], image segmentation [30], [35], [68], video analysis and detection [18], [74], [83], image retrieval [46], image co-segmentation [15], [16], [85]. Most of the existing Salient Object Detection (SOD) methods [33], [48], [81] mainly deal with RGB images. However, they usually produce inaccurate SOD results on the scenarios of similar texture, complex background, or homogeneous objects [73], [82].

With the popularity of depth sensors in smartphones, depth maps associated with the corresponding RGB images are becoming much easier to acquire. Intuitively, the depth information, *e.g.*, 3D layout and spatial cues, is crucial for reducing the ambiguity in the RGB images, and serves as important supplements to improve the SOD performance [37]. Thus, RGB-D SOD has received increasing research attention [7], [19], [23], [24], [61], [78]–[80].

For current RGB-D SOD methods, the depth contrast has served as the most important prior [59], [64], [66], [86], and it is often used to shift more priority on the foreground regions which have a strong contrast with the background. For example, among early RGB-D SOD works, Fan *et al.* [20] employ the depth map as the weighting factor for color contrast. The recent work of CPFP [86] designs an effectiveness loss to

Manuscript received April 26, 2020; revised November 8, 2020 and December 16, 2020; accepted December 23, 2020. Date of publication January 13, 2021; date of current version January 20, 2021. This work was supported in part by the Major Project for New Generation of AI under Grant 2018AAA0100400, Grant NSFC 61922046, and Grant NSFC 61972216; in part by the National Youth Talent Support Program; and in part by the Tianjin Natural Science Foundation under Grant 17JCJQJC43700 and Grant 18JCYBJC41300. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Charith Abhayaratne. *(Corresponding author: Shao-Ping Lu.)*

Zhao Zhang, Zheng Lin, Jun Xu, Shao-Ping Lu, and Deng-Ping Fan are with the TKLNDST, College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: slu@nankai.edu.cn).

Wen-Da Jin is with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China.

Digital Object Identifier 10.1109/TIP.2021.3049959

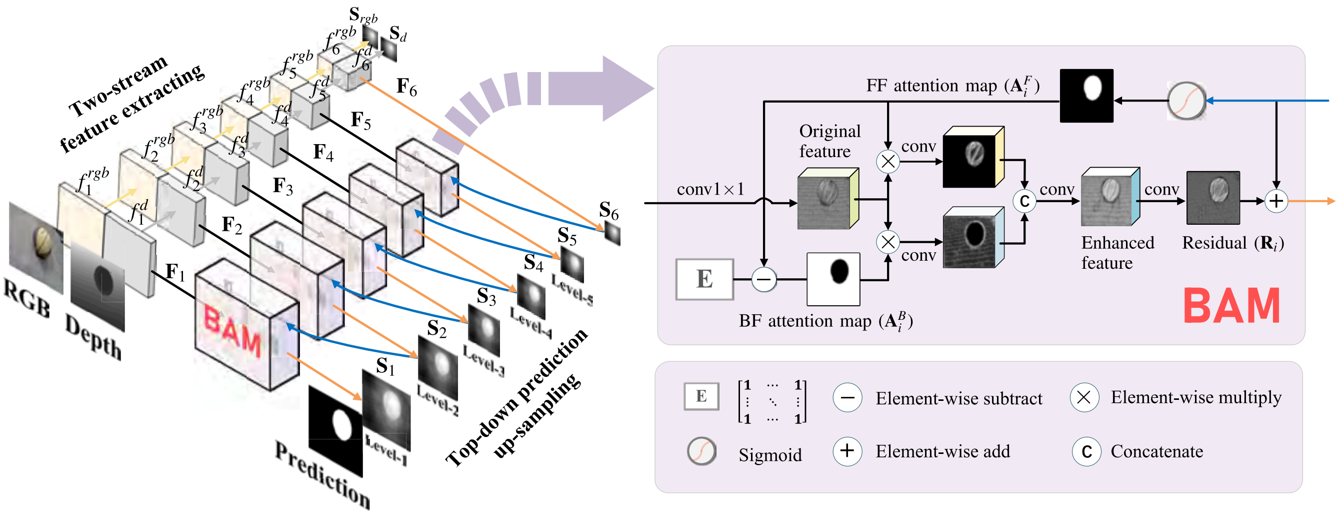

Fig. 2. **The overall architecture of our BiANet.** BAM denotes the proposed Bilateral Attention Module, and it also can be selectively replaced by its multi-scale extension (MBAM). BiANet contains three main steps: two-stream feature extracting, top-down prediction up-sampling, and bilateral attention residual compensation (by BAM). Specifically, it first extracts the multi-level features $\{f_i^{rgb}, f_i^d\}_{i=1}^6$ from the RGB and depth streams, and concatenates them to $\{\mathbf{F}_i\}_{i=1}^6$. The salient maps of $\mathbf{S}_{rgb}$ and $\mathbf{S}_d$ are predicted from f_6^{rgb} and f_6^d for deep supervision. We take the top feature $\mathbf{F}_6$ to predicate a coarse salient map $\mathbf{S}_6$. To obtain the accurate and high-resolution result, we up-sample the initial salient map and compensate the details by BAMs in a top-down manner. BAMs receive the higher-level prediction $\mathbf{S}_{i+1}$ and current level feature $\mathbf{F}_i$ as inputs. In a BAM, the foreground-first attention map $\mathbf{A}_i^F$ and the background-first attention map $\mathbf{A}_i^B$ can be calculated according to $\mathbf{S}_{i+1}$. We apply the dual complementary attention maps to explore the foreground and background cues bilaterally, and jointly infer the residual for refining the up-sampled saliency map.

enhance the depth contrast for better inducing the network to focus on the foreground regions. More attention on the foreground region is indeed conducive to learning salient cues. Meanwhile, as demonstrated in [44], [75], [76], understanding background information in a scene can help promote the SOD performance. The foreground and background priors are largely different. For example, the foreground priors contain more cues that attract human visual attention, such as sensitive categories, bright colors, special shapes, closer distance to the observer, while the background (non-salient) priors are the opposite. Therefore, it is necessary to explore the foreground and background cues respectively and then jointly to mine the accurate salient region in a scene. Several traditional methods [38], [38], [77] have predicted salient objects in this way. Benefiting from together exploring the foreground and background cues, these methods achieved the leading effect at that time. However, this simple and effective idea is largely ignored by current RGB-D SOD networks.

In this paper, we propose a Bilateral Attention Network (BiANet) to collaboratively learn complementary foreground and background features from both RGB and depth streams for better RGB-D SOD performance. As shown in Figure 2, our BiANet employs a two-stream architecture, and the side outputs from the RGB and depth streams are concatenated in multiple stages. Firstly, we use the high-level semantic features $\mathbf{F}_6$ to locate the foreground and background regions $\mathbf{S}_6$. However, the initial saliency map $\mathbf{S}_6$ is coarse and in low-resolution. To enhance the coarse saliency map, we design a Bilateral Attention Module (BAM), which is composed of the complementary foreground-first (FF) attention and background-first (BF) attention mechanisms. The FF shifts attention on the foreground region to gradually refine its saliency prediction, while the BF focuses on the background region to recover the potential salient regions around the boundaries. By bilaterally exploring the foreground and background cues, the model helps predict more accurately, as shown in Figure 1. Secondly, we propose a multi-scale extension of BAM (MBAM) to effectively learn multi-scale contextual information, and capture both local and global saliency information to further improve the SOD performance. Extensive experiments on six benchmark datasets demonstrate that our BiANet achieves better performance than previous state-of-the-arts on RGB-D SOD, and is very fast owing to our simple architecture.

In summary, our main contributions are three-fold:

- **We propose a simple yet effective Bilateral Attention Module (BAM)** to explore the foreground and background cues collaboratively with the rich foreground and background information from the depth images.
- **Our BiANet achieves better performance on six popular RGB-D SOD datasets** under nine standard metrics, and presents better visual effects (*e.g.*, contains more details and sharp edges) than the state-of-the-art methods.
- **Our BiANet runs at 34~80 *fps*** on an NVIDIA GeForce RTX2080Ti GPU under different settings, and is a feasible solution for real-world applications.

The remainder of this paper is organized as follows. In §II, we briefly survey the related work. In §III, we present the proposed Bilateral Attention Network (BiANet) for RGB-D salient object detection. Extensive experiments are conducted in §IV to evaluate its performance when compared with state-of-the-art RGB-D SOD methods on six benchmark datasets. The conclusion is given in §V.

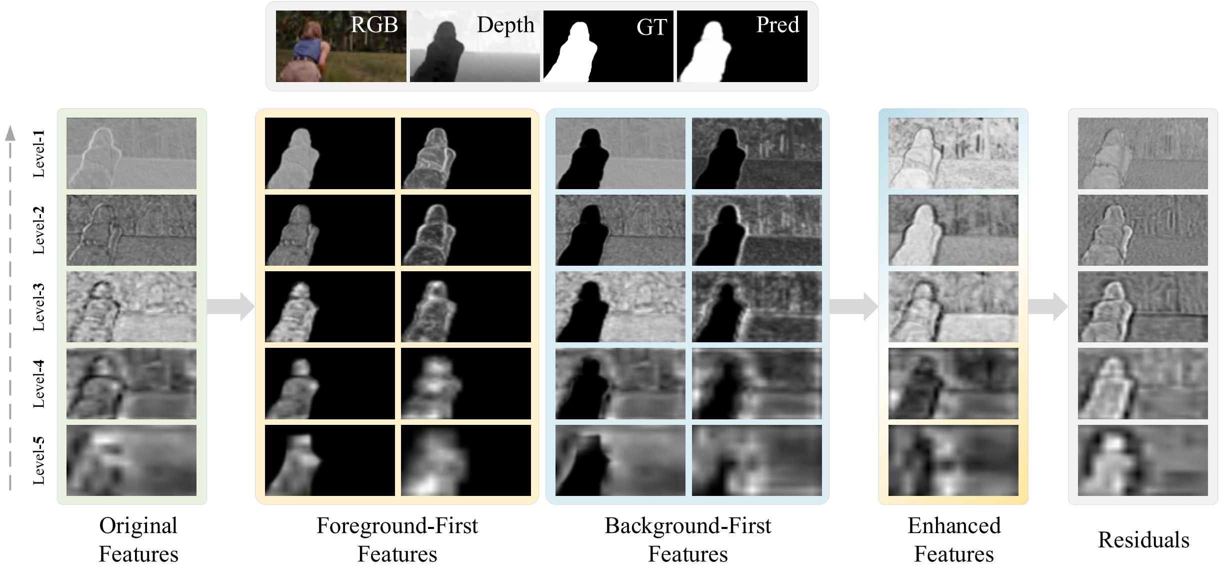

Fig. 3. **Visualizing the working mechanism of bilateral attention**. The original features are the averaged side-output features in each level. We show the original features directly multiplied by foreground- and background-first attention maps in left columns of yellow and blue boxes. The right columns of the two boxes are the further convoluted features in two branches. As can be seen, the foreground-first features focus on foreground region to explore the saliency cues; while the background-first features shift more attention to the background regions to mine the potentially significant objects. No matter in the features of foreground- or background-first features, more priority is shifted to the uncertain (low confidence) areas caused by the up-sampling. When fusing the two branches and jointly inferring, we can see the bilaterally enhanced features have a more accurate understanding where the foreground or background is. Due to obtaining more attention, the uncertain areas are reassigned to the right attribution by the residual with strong contrast. 'Pred' is the prediction of the model.

II. Related Work

A. RGB-D Salient Object Detection

RGB-D salient object detection (SOD) aims to segment the most attractive object(s) in a pair of cross-modal RGB and depth images. Early methods mainly focus on extracting low-level saliency cues from RGB and depth images, exploring object distance [37], difference of Gaussian [34], graph knowledge [12], multi-level discriminative saliency fusion [66], multi-contextual contrast [11], [59], background enclosure [21], *etc.*. However, these methods easily result in inaccurate saliency predictions due to the lack of high-level feature representation. Recently, Qu *et al.* [63] introduce the deep neural networks (DNNs) to investigate high-level representations of multiple saliency cues, including local and global contrast, and color compactness. After that, DNNs have been largely employed to find the high-level representations of RGB and depth images in this task [7], [23], [39], [79]. For instance, some works [8], [27], [70] first extract the RGB and depth features separately and then fuse them in the shallow, middle, or deep layers of the network. The methods of [6], [7], [40], [61] further improve the SOD performance by fusing cross-modal features in multi-level stages instead of as a one-off integration. Fan *et al.* [17] propose that the depth maps are not always beneficial to salient object detection; thus, they propose a depth depurator unit to automatically discard some low-quality depth maps.

B. Foreground and Background Cues

There are great differences in the distribution of foreground and background, so it is necessary to explore their respective cues. In traditional methods, some works focus on reasoning salient areas in foreground and background jointly. Yang *et al.* [77] proposed a two-stage method for SOD. It first regards the top, bottom, left, and right marginal regions of the input images as background seeds to infer the possible foreground super-pixels via a graph-based manifold ranking. Then, it ranks the graph for final prediction depending on the foreground seeds. Ren *et al.* [64] adopt boundary connectivity to locate the initial background regions instead of only assuming the boundaries as background. Liang *et al.* [44] introduce the depth map to take the region that far away from the observer as the initial background region.

III. Proposed BiANet for RGB-D SOD

In this section, we first introduce the overall architecture of our BiANet, and then present the bilateral attention module (BAM) as well as its multi-scale extension (MBAM).

A. Architecture Overview

As shown in Figure 2, our Bilateral Attention Network (BiANet) contains three main steps: feature extracting, prediction up-sampling, and bilateral attention residual compensation. We extract the multi-level features from the RGB and depth streams. With increasing network depth, the high-level features (*e.g.*, $\mathbf{F}_4$) will be more potent for capturing global context, while it loses the object details. When we up-sample the high-level predictions, the saliency maps (*e.g.*, $\mathbf{S}_5$) will be blurred, and the edges will become difficult to find. That is, the prediction value of the pixel location is near 0.5 after the Sigmoid layer. Thus, we use the proposed Bilateral Attention

Module (BAM) to distinguish these uncertain regions to the foreground or background.

1) Feature Extracting: We encode RGB and depth information with two streams. Specifically, both the RGB and depth streams employ five convolutional blocks from VGG-16 [65] as the standard backbone and attach an additional convolutional group with three convolutional layers to predict the saliency maps, respectively. Unlike previous works [8], [27], [90], we explore the cross-modal fusion of RGB and depth features at multiple stages, rather than fusing them once in low or high stage. The i-th side output f_i^{rgb} from the RGB stream and f_i^d from the depth stream are concatenated as a feature tensor $\mathbf{F}_i$. Note that $\mathbf{F}_6$ is concatenated by $M(f_5^{rgb})$ and $M(f_5^d)$, where $M(\cdot)$ denotes the max-pooling operation. The coarse saliency map $\mathbf{S}_6$ is predicted from $\mathbf{F_6}$ using two 3×3 convolutional layers, and $\{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_5\}$ are prepared for the BAMs in our BiANet to further refine the up-sampled saliency maps, by distinguishing the uncertain regions as foreground or background in a top-down manner.

2) Prediction up-Sampling: The initial saliency map predicted from the high-level features is coarse in low-resolution, but useful to predict the initial position of the foreground and background, since it contains rich semantic information. To refine the basic saliency map $\mathbf{S}_6$, a lower-level feature $\mathbf{F}_5$ with more details is used to predict the residual component between the higher-level prediction and the ground-truth (**GT**) with the help of BAM. We add the predicted residual component $\mathbf{R}_5$ to the up-sampled higher-level prediction S_6, and obtain a refined prediction $\mathbf{S}_5$, *etc.*, that is,

$$\mathbf{S}_i = \mathbf{R}_i + U(\mathbf{S}_{i+1}), \quad i \in \{1, \ldots, 5\}, \tag{1}$$

where $U(\cdot)$ means up-sampling. Finally, our BiANet obtains a saliency map by $\mathbf{S} = \sigma(\mathbf{S}_1)$, where $\sigma(\cdot)$ is a Sigmoid function.

3) Bilateral Attention Residual Compensation: To get better residuals and distinguish up-sampled foreground and background regions, we design a bilateral attention module (BAM) to enable our BiANet to discriminate the foreground and background. In our BAM, the higher-level prediction serves as a foreground-first attention (FF) map, and the reversed prediction serves as background-first (BF) attention map to combine the bilateral attention on foreground and background. In Figure 3, one can see that the residual generated by BAM possesses high contrast at the object boundaries. More details are described in Sections III-B and III-C.

4) Loss Function: Deep supervision is widely used in the SOD task [22], [29]. It clarifies the optimization goals for each step of the network, and accelerates the convergence of training. For quick convergence, we also apply deep supervision in the depth stream output $\mathbf{S}_d$, RGB stream output $\mathbf{S}_{rgb}$, and each top-down side output $\{\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_6\}$. The total loss function of our BiANet is

$$\mathcal{L} = \sum\nolimits_{i=1}^{6} w_i \mathcal{L}_{ce}\left(\sigma\left(\mathbf{S}_i\right), \mathbf{GT}\right) + w_d \mathcal{L}_{ce}\left(\sigma\left(\mathbf{S}_d\right), \mathbf{GT}\right) + w_{rgb} \mathcal{L}_{ce}\left(\sigma\left(\mathbf{S}_{rgb}\right), \mathbf{GT}\right), \tag{2}$$

in which w_i, w_d, and w_{rgb} are the weight coefficients and simply set to 1 in our experiments. $\mathcal{L}_{ce}(\cdot)$ is the binary cross entropy loss, which is formulated as

$$\mathcal{L}_{ce}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{N}\sum_{i=1}^{N}\Big(y_i log(x_i) + (1 - y_i) log(1 - x_i)\Big). \tag{3}$$

In the above equation, $x_i \in \mathbf{X}$ and $y_i \in \mathbf{Y}$, and N denotes the total pixel number.

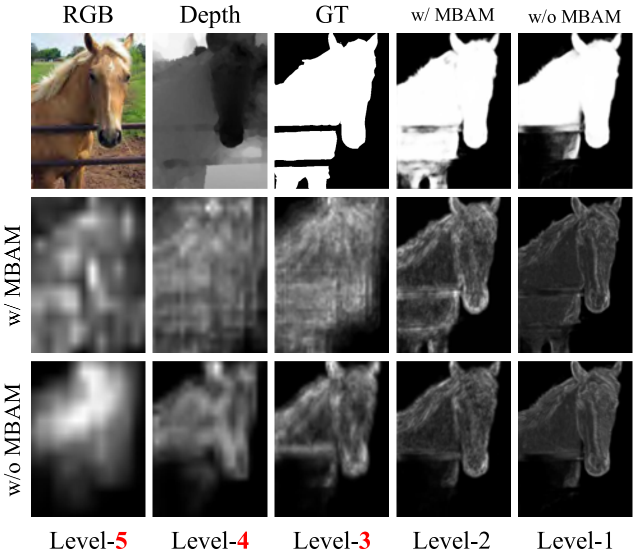

Fig. 4. **Comparison of the high-level features captured by MBAM and BAM.** The second row is the averaged foreground-first features from the model where the MBAMs are applied in the top three levels (marked with red numbers). The third row is the averaged foreground-first features obtained from the model in which all levels are armed with BAMs. We can see that, compared with applying the BAMs, MBAMs in higher levels capture more complete information, which is conducive to the object locating as shown in the first row.

B. Bilateral Attention Module (BAM)

Given the initial foreground and background, how to refine the prediction using higher-resolution cross-modal features is the focus of this paper. Considering that the distribution of foreground and background are quite different, we design a bilateral attention module using a pair of reversed attention components to learn features from the foreground and background respectively, and then jointly refine the prediction. As can be seen in Figure 2, to focus more on the foreground, we use the up-sampled prediction from the higher-level as foreground-first attention (FF) maps $\{\mathbf{A}^F\}_{i=1}^5$ after they are activated by Sigmoid function, and the background-first attention (BF) maps $\{\mathbf{A}^B\}_{i=1}^5$ are generated by subtracting FF maps from matrix $\mathbf{E}$, in which all the elements are 1. That is,

$$\begin{cases} \mathbf{A}_i^F = \sigma\Big(U(\mathbf{S}_{i+1})\Big), \\ \mathbf{A}_i^B = \mathbf{E} - \sigma\Big(U(\mathbf{S}_{i+1})\Big), \end{cases} \quad i \in \{1, 2, 3, 4, 5\}. \tag{4}$$

Then, as shown in Figure 2, we apply FF and BF to weight the side-output features in two branches, respectively, and further predict the residual component jointly.

$$\mathbf{R}_i = \mathcal{P}_i^R\left(\left[\mathcal{P}_i^F\left(\hat{\mathbf{F}}_i \odot \mathbf{A}_i^F\right), \mathcal{P}_i^B\left(\hat{\mathbf{F}}_i \odot \mathbf{A}_i^B\right)\right]\right). \tag{5}$$

In the above equation, $\odot$ denotes element-wise multiplication. $\hat{\mathbf{F}}_i$ is the channel-reduced feature of $\mathbf{F}_i$ using 32 1×1 convolutions to reduce the computational cost. $\mathcal{P}_i^F$ and $\mathcal{P}_i^B$ are the foreground-first and background-first branches. They consist of 32 convolutional kernels with a size of 3×3 and a ReLU layer. The $[\mathbf{X}, \mathbf{Y}]$ means concatenating the $\mathbf{X}$ and $\mathbf{Y}$ in channel-wise. $\mathcal{P}_i^R$ is the prediction layer to output a single channel residual map via 3×3 kernels basing on the concatenated features. Once the $\mathbf{R}_i$ is obtained, the refined prediction $\mathbf{S}_i$ is obtained via Equation 1.

To better understand the working mechanism of BAM, in Figure 3, we visualize the channel-wise averaged features from BAMs in different levels. In BAM, the original features will be first fed into two branches by multiply the FF and BF attention maps, respectively. The result of the direct multiplication is illustrated in the left half of the yellow (FF features) and blue (BF features) boxes. We can see that FF branch shifts attention to the foreground area predicted from its higher level to explore foreground saliency cues. After a convolutional layer, more priority is given to the uncertain area. Complementarily, BF branch focuses on the background area to explore the background cues, looking for possible salient objects within it. In our BiANet, the top-down prediction up-sampling is a process in which the resolution of salient objects is gradually increased. It will result in uncertain coarse boundaries. Improving the quality of object edge segmentation is important for segmentation tasks. Many methods, based on active contours [51], [52], [54]–[57], boundary supervision [42], [67], [88], *etc.*, are proposed to shift more attention on boundary regions. That is coincident with the original intention and advantages of our cascaded bilateral attention model. We can see that both FF and BF features focus on these boundaries. The low-level and high-resolution FF branch will eliminate the overflow of the uncertain area, while the BF branch will eliminate the uncertain area which does not belong to the background. That is an important reason why BiANet performs better on detail and is prone to predicting sharp edges. After the joint inferring, we can see the bilaterally enhanced features contain more discriminative spatial information of foreground and background. The generated residual components are with sharp contrast on the edges, and then suppress the background area and strengthen the foreground regions.

C. Multi-Scale Extension of BAM (MBAM)

Salient objects in a scene are various in location, size, and shape. Thus, exploring the multi-scale context in high-level layers benefits for understanding the scene [71], [87]. To this end, we extend our BAM with a multi-scale version, in which groups of dilated convolutions are used to extract pyramid representations from the undetermined foreground and background areas. Specifically, the module can be described as

$$\mathbf{R}_i = \mathcal{P}_i^R\left(\left[\sqcup_{j=1}^{4}\mathcal{D}_{ij}^F\left(\mathbf{F}_i \odot \mathbf{A}_i^F\right), \sqcup_{j=1}^{4}\mathcal{D}_{ij}^B\left(\mathbf{F}_i \odot \mathbf{A}_i^B\right)\right]\right), \tag{6}$$

where $\sqcup$ means a concatenate operation. $\mathcal{D}_{i1}^F$ and $\mathcal{D}_{i1}^B$ consist of 1×1 kernels with 32 channels and a ReLU layer. $\{\mathcal{D}_{ij}^F\}_{j=2}^4$ and $\{\mathcal{D}_{ij}^B\}_{j=2}^4$ are groups of dilated convolutions, with rates of 3, 5, and 7. They all consist of 3×3 kernels with 32 channels and a ReLU layer.

We recommend applying the MBAM in high-level cross-modal features, such as $\{\mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5\}$, which need different sizes of receptive fields to explore multi-scale context. MBAM effectively improves the detection performance but introduces a certain computational cost. Thus, the number of MBAM should be a trade-off in practical applications. In Section IV-C.3, we discuss in detail how the number of MBAM changes the detection effect and calculation cost.

In order to intuitively observe the gain effect brought by MBAM, we visualize the averaged foreground-first feature maps from MBAMs and BAMs in Figure 4. In the second row, the feature maps are obtained from the model with three MBAMs in its top three levels, while in the last row, all the feature maps are collected from BAMs. We can see the target object (horse) account for a large proportion of the scene. Without the ability to perceive multi-scale information effectively, the BAM fails to capture the accurate global salient regions in high levels and leads to incomplete prediction finally. When introducing the multi-scale extension, we can see higher-level features achieve stronger spatial representation, supporting locating a more complete salient object.

D. Implementation Details

1) Settings: We apply the MBAM in the high-level side outputs $\{\mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5\}$ during implementation, and use bilinear interpolation in all interpolation operations. The initial parameters of our backbone are loaded from a VGG-16 network pre-trained on ImageNet. Our BiANet is based on PyTorch [58].

2) Training: Following D3Net [17], we use the training set containing 1485 and 700 image pairs from the *NJU2K* [34] and *NLPR* [59] datasets, respectively. The rest samples of *NJU2K* are used as the validation set. We employ the Adam optimizer [36] with an initial learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. The batch size is set to 8. We train our BiANet until 30 epochs for VGG backbone and 50 epochs for ResNet and Res2Net backbone. The training images are resized to 224×224, also during the test. The output saliency maps are resized back to the original size for evaluation. Accelerated by an NVIDIA GeForce RTX 2080Ti, our BiANet (VGG-16 backbone) is trained for less than three hours and runs at 34∼80 *fps* (with different numbers of MBAMs) for the inputs with 224×224 resolution.

IV. Experiments

A. Evaluation Protocols

1) Evaluation Datasets: We conduct experiments on six widely used RGB-D based SOD datasets. *NJU2K* [34] and *NLPR* [59] are two popular large-scale RGB-D SOD datasets containing 1985 and 1000 images, respectively. *DES* [9] contains 135 indoor images with fine structures collected with Microsoft Kinect [84]. *STERE* [50] contains 1000 internet images, and the corresponding depth maps are generated by stereo images using a sift flow algorithm [45]. *SSD* [91] is a small-scale but high-resolution dataset with 400 images in

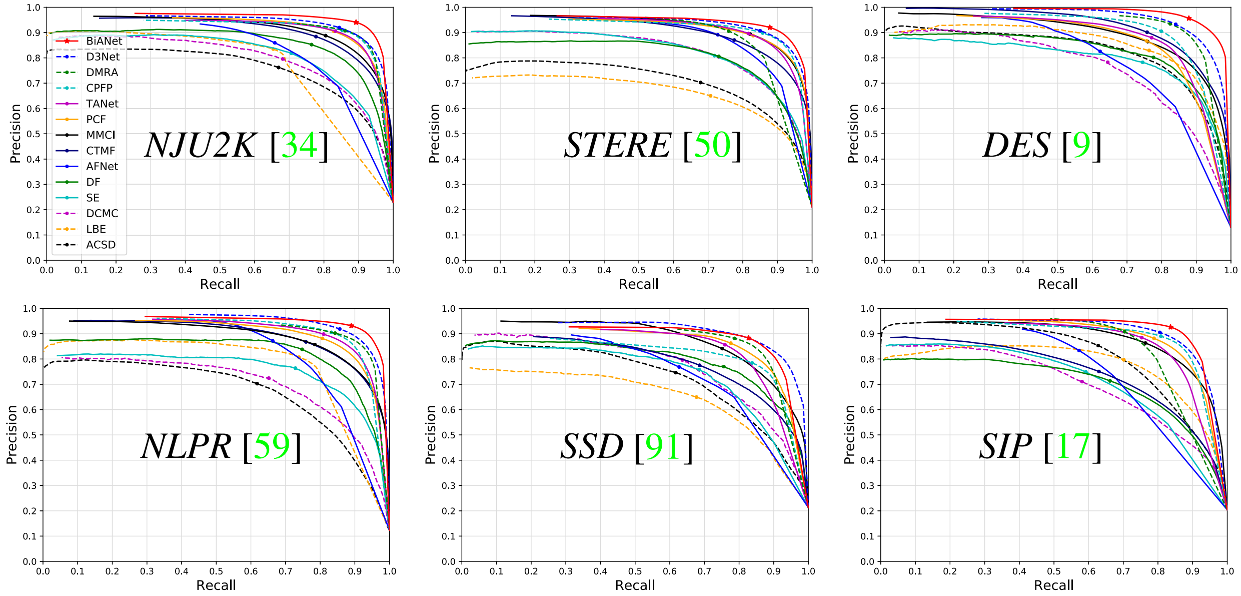

Fig. 5. **PR curves of our BiANet and other 13 state-of-the-art methods across 6 datasets**. The node on each curve denotes the precision and recall value used for calculating maximal F-measure.

960 × 1080 resolution. *SIP* [17] is a high-quality RGB-D SOD dataset with 929 person images.

2) Evaluation Metrics: We employ nine metrics to comprehensively evaluate these methods. **Precision-Recall (PR) curve** [62] shows the precision and recall performances of the predicted saliency map at different binary thresholds. **F-measure (F_β)** [1] is computed by the weighted harmonic mean of the thresholded precision and recall. We employ maximum F-measure as suggest in [4]. **Mean Absolute Error** (MAE, $\mathcal{M}$) [60] directly estimates the average pixel-wise absolute difference between the prediction and the binary ground-truth map. **S-measure (S_α)** [13] is an advanced metric, which takes the region-aware and object-aware structural similarity into consideration. **E-measure (E_ξ)** [14] is the recent proposed Enhanced alignment measure in the binary map evaluation field, which combines local pixel values with the image level mean value in one term, jointly capturing image-level statistics and local pixel matching information. We take the maximum E-measure following [5], [40]. **Mean-Square Error (MSE)** measures the average of the squares of the errors. **Peak Signal-to-Noise Ratio (PSNR)** is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. The higher PSNR, the better the quality of prediction. **Structural Similarity (SSIM)** evaluates the similarity of the two images in terms of luminance, contrast, and structure. The metrics of MSE, PSNR, SSIM are widely used in watermarking [2], [3], image compression [72], image enhancement [43], *etc.*.

B. Comparison With State-of-the-Arts

1) Comparison Methods: We compared with 13 state-of-the-art RGB-D SOD methods, including 4 traditional methods: ACSD [34], LBE [21], DCMC [12], MDSF [66], and SE [26], and 9 DNN-based methods: DF [63], AFNet [70], CTMF [27], MMCI [8], PCF [6], TANet [7], CPFP [86], DMRA [61], and D3Net [17]. The codes and saliency maps of these methods are provided by the authors.

2) Quantitative Evaluation: The complete quantitative evaluation results are listed in Table I. The comparison methods are presented from right to left according to the comprehensive performance of these metrics, where the lower the value of MAE ($\mathcal{M}$), the better the model's effect. The other metrics are the opposite. We also plot the PR curves of these methods in Figure 5. One can see that our BiANet achieves remarkable advantages over the comparison methods. DMRA [61] and D3Net [17] are well-matched in these datasets. On the large-scaled *NJU2K* [34] and *NLPR* [59] datasets, our BiANet outperforms the second best with ∼3% improvement on F_β. On the *DES* [9] dataset, Compared to methods that are heavily dependent on depth information, our proposed BiANet also has a 3.8% improvement on F_β. This indicates that our BiANet can make more efficient use of depth information. Although the *SSD* [91] dataset is high-resolution, the quality of the depth map is poor. Our BiANet still exceeds D3Net [17], which is specifically designed for robustness to low-quality depth maps. Our BiANet also performs the best on the *SIP* [17], which is a challenging dataset with complex scenes and multiple objects.

3) Qualitative Results: To further demonstrate the effectiveness of our BiANet, we visualize the saliency maps of our BiANet and other five methods with highest quantitative results in Figure 6. One can see that the target object in the 1st column is tiny, and its white shoes and hat are hard to distinguish from the background. Our BiANet effectively utilizes the depth information, while the others are disturbed by RGB background clutter. The inputs in the 2nd column are challenging because the depth map is mislabeled, and the RGB image was taken in a dark environment with low

TABLE I

QUANTITATIVE COMPARISONS OF OUR BIANET WITH NINE DEEP-LEARNING-BASED METHODS AND FOUR TRADITIONAL METHODS ON SIX POPULAR DATASETS IN TERM OF S-MEASURE (S_α), MAXIMUM F-MEASURE (F_β), MAXIMUM E-MEASURE (E_ξ), MEAN ABSOLUTE ERROR (MAE, $\mathcal{M}$), MEAN-SQUARE ERROR (MSE), PEAK SIGNAL-TO-NOISE RATIO (PSNR), AND STRUCTURAL SIMILARITY (SSIM). ↑ MEANS THAT THE LARGER THE NUMERICAL VALUE, THE BETTER THE MODEL, WHILE ↓ MEANS THE OPPOSITE. FOR TRADITIONAL METHODS, THE STATISTICS ARE BASED ON OVERALL DATASETS RATHER ON THE TEST SET. WE SHOW THE PERFORMANCES OF OUR BIANET BASED ON DIFFERENT BACKBONES. VGG11 AND VGG16 IS THE VGG NETWORK PROPOSED IN [65]. RES50 IS RESNET-50 PROPOSED IN [28]. RES^{2}50 IS RES2NET-50 PROPOSED IN [25]. THE BIANET WITH VGG-16 BACKBONE IS USED FOR COMPARISON. WE USE **BOLD** TO HIGHLIGHT THE BEST RESULTS AND UNDERLINE TO HIGHLIGHT THE SECOND BEST RESULTS

		ACSD	LBE	DCMC	SE	DF	AFNet	CTMF	MMCI	PCF	TANet	CPFP	DMRA	D3Net	BiANet (Ours)			
	Metric	[34]	[21]	[12]	[26]	[63]	[70]	[27]	[8]	[6]	[7]	[86]	[61]	[17]	vgg16	vgg11	Res50	Res^{2}50
NJU2K [34]	S_α ↑	0.699	0.695	0.686	0.664	0.763	0.772	0.849	0.858	0.877	0.878	0.879	0.886	$\underline{0.893}$	**0.915**	0.912	0.917	0.923
	F_β ↑	0.711	0.748	0.715	0.748	0.804	0.775	0.845	0.852	0.872	0.874	0.877	0.886	$\underline{0.887}$	**0.920**	0.913	0.920	0.925
	E_ξ ↑	0.803	0.803	0.799	0.813	0.864	0.853	0.913	0.915	0.924	0.925	0.926	0.927	$\underline{0.930}$	**0.948**	0.947	0.949	0.952
	$\mathcal{M}$ ↓	0.202	0.153	0.172	0.169	0.141	0.100	0.085	0.079	0.059	0.060	0.053	$\underline{0.051}$	$\underline{0.051}$	**0.039**	0.040	0.036	0.034
	MSE↓	0.105	0.117	0.106	0.127	0.079	0.087	0.045	0.044	0.039	0.041	0.041	0.043	$\underline{0.035}$	**0.030**	0.030	0.029	0.027
	PSNR↑	10.76	11.13	11.09	10.84	12.67	12.55	14.75	15.20	16.44	16.33	16.60	16.93	$\underline{17.22}$	**18.96**	18.71	19.14	19.48
	SSIM↑	0.336	0.811	0.512	0.691	0.546	0.822	0.689	0.699	0.822	0.832	0.891	$\underline{0.903}$	0.866	**0.913**	0.909	0.923	0.926
STERE [50]	S_α ↑	0.692	0.660	0.731	0.708	0.757	0.825	0.848	0.873	0.875	0.871	0.879	0.835	$\underline{0.889}$	**0.904**	0.899	0.905	0.908
	F_β ↑	0.669	0.633	0.740	0.755	0.757	0.823	0.831	0.863	0.860	0.861	0.874	0.847	$\underline{0.878}$	**0.898**	0.892	0.899	0.904
	E_ξ ↑	0.806	0.787	0.819	0.846	0.847	0.887	0.912	0.927	0.925	0.923	0.925	0.911	$\underline{0.929}$	**0.942**	0.941	0.943	0.942
	$\mathcal{M}$ ↓	0.200	0.250	0.148	0.143	0.141	0.075	0.086	0.068	0.064	0.060	$\underline{0.051}$	0.066	0.054	**0.043**	0.045	0.040	0.039
	MSE↓	0.099	0.117	0.084	0.101	0.078	0.062	0.046	0.038	0.040	0.041	0.041	0.057	$\underline{0.037}$	**0.032**	0.034	0.032	0.031
	PSNR↑	10.67	9.65	11.97	11.57	12.51	13.97	14.40	15.73	15.77	15.54	16.26	14.39	$\underline{16.71}$	**17.78**	17.21	17.85	18.05
	SSIM↑	0.318	0.213	0.523	0.668	0.487	0.849	0.682	0.739	0.801	0.837	$\underline{0.894}$	0.885	0.850	**0.902**	0.898	0.915	0.918
DES [9]	S_α ↑	0.728	0.703	0.707	0.741	0.752	0.770	0.863	0.848	0.842	0.858	0.872	$\underline{0.900}$	0.898	**0.931**	0.943	0.930	0.942
	F_β ↑	0.756	0.788	0.666	0.741	0.766	0.728	0.844	0.822	0.804	0.827	0.846	$\underline{0.888}$	0.880	**0.926**	0.938	0.927	0.942
	E_ξ ↑	0.850	0.890	0.773	0.856	0.870	0.881	0.932	0.928	0.893	0.910	0.923	$\underline{0.943}$	0.935	**0.971**	0.979	0.968	0.978
	$\mathcal{M}$ ↓	0.169	0.208	0.111	0.090	0.093	0.068	0.055	0.065	0.049	0.046	0.038	$\underline{0.030}$	0.033	**0.021**	0.019	0.021	0.017
	MSE↓	0.058	0.071	0.058	0.058	0.053	0.058	0.029	0.033	0.035	0.032	0.029	0.025	$\underline{0.021}$	**0.014**	0.012	0.015	0.013
	PSNR↑	12.74	11.94	12.85	13.70	13.85	14.08	16.52	16.14	16.85	17.03	17.96	18.77	$\underline{19.17}$	**20.50**	20.61	20.05	20.59
	SSIM↑	0.181	0.134	0.505	0.700	0.557	0.866	0.774	0.655	0.871	0.885	0.919	$\underline{0.937}$	0.901	**0.943**	0.943	0.947	0.951
NLPR [59]	S_α ↑	0.673	0.762	0.724	0.756	0.802	0.799	0.860	0.856	0.874	0.886	0.888	0.899	$\underline{0.905}$	**0.925**	0.927	0.926	0.929
	F_β ↑	0.607	0.745	0.648	0.713	0.778	0.771	0.825	0.815	0.841	0.863	0.867	0.879	$\underline{0.885}$	**0.914**	0.914	0.917	0.919
	E_ξ ↑	0.780	0.855	0.793	0.847	0.880	0.879	0.929	0.913	0.925	0.941	0.932	$\underline{0.947}$	0.945	**0.961**	0.962	0.962	0.963
	$\mathcal{M}$ ↓	0.179	0.081	0.117	0.091	0.085	0.058	0.056	0.059	0.044	0.041	0.036	$\underline{0.031}$	0.033	**0.025**	0.024	0.023	0.023
	MSE↓	0.069	0.053	0.061	0.057	0.041	0.049	0.029	0.032	0.029	0.027	0.028	0.026	$\underline{0.022}$	**0.018**	0.018	0.018	0.018
	PSNR↑	12.61	15.48	13.84	15.09	16.18	15.53	16.97	16.82	18.07	18.41	19.26	19.17	$\underline{19.61}$	**21.10**	21.00	21.14	21.21
	SSIM↑	0.248	0.896	0.544	0.743	0.626	0.881	0.770	0.730	0.869	0.881	0.922	$\underline{0.933}$	0.901	**0.941**	0.941	0.948	0.949
SSD [91]	S_α ↑	0.675	0.621	0.704	0.675	0.747	0.714	0.776	0.813	0.841	0.839	0.807	0.857	$\underline{0.865}$	**0.867**	0.861	0.863	0.863
	F_β ↑	0.682	0.619	0.711	0.710	0.735	0.687	0.729	0.781	0.807	0.810	0.766	0.844	$\underline{0.846}$	**0.849**	0.839	0.843	0.843
	E_ξ ↑	0.785	0.736	0.786	0.800	0.828	0.807	0.865	0.882	0.894	0.897	0.852	0.906	$\underline{0.907}$	**0.916**	0.899	0.911	0.901
	$\mathcal{M}$ ↓	0.203	0.278	0.169	0.165	0.142	0.118	0.099	0.082	0.062	0.063	0.082	$\underline{0.058}$	0.059	**0.051**	0.054	0.048	0.050
	MSE↓	0.107	0.138	0.102	0.128	0.089	0.104	0.066	0.049	0.042	0.044	0.069	0.050	**0.040**	**0.040**	0.043	0.040	0.042
	PSNR↑	10.61	9.44	11.61	11.18	12.55	12.01	13.22	14.84	16.22	15.94	14.96	15.95	$\underline{16.68}$	**17.72**	17.34	17.49	17.62
	SSIM↑	0.257	0.195	0.491	0.663	0.542	0.811	0.706	0.732	0.846	0.850	0.861	$\underline{0.900}$	0.865	**0.902**	0.894	0.914	0.911
SIP [17]	S_α ↑	0.732	0.727	0.683	0.628	0.653	0.720	0.716	0.833	0.842	0.835	0.850	0.806	$\underline{0.864}$	**0.883**	0.877	0.887	0.889
	F_β ↑	0.763	0.751	0.618	0.661	0.657	0.712	0.694	0.818	0.838	0.830	0.851	0.821	$\underline{0.861}$	**0.890**	0.882	0.890	0.893
	E_ξ ↑	0.838	0.853	0.743	0.771	0.759	0.819	0.829	0.897	0.901	0.895	0.903	0.875	$\underline{0.910}$	**0.925**	0.924	0.926	0.928
	$\mathcal{M}$ ↓	0.172	0.200	0.186	0.164	0.185	0.118	0.139	0.086	0.071	0.075	0.064	0.085	$\underline{0.063}$	**0.052**	0.054	0.047	0.047
	MSE↓	0.093	0.083	0.107	0.137	0.121	0.107	0.098	0.055	0.053	0.058	0.055	0.078	$\underline{0.048}$	**0.043**	0.044	0.040	0.040
	PSNR↑	11.12	11.38	10.56	10.13	10.35	11.37	11.32	14.13	14.83	14.47	15.04	13.66	$\underline{15.56}$	**17.14**	16.61	17.33	17.47
	SSIM↑	0.454	0.285	0.412	0.706	0.459	0.816	0.666	0.738	0.838	0.834	$\underline{0.892}$	0.874	0.859	**0.906**	0.900	0.918	0.918

contrast. Our BiANet successfully detects the target sculpture and eliminates the interference of flowers and the base of the sculpture, while D3Net mistakenly detects a closer rosette, and DMRA loses the part of the object that is similar to the background. The 3rd column shows the ability of our BiANet to detect complex structures and textureless [53] of salient objects. Among these methods, only our BiANet completely discover the chairs, including the fine legs. The 4th column is a multi-object scene. Because there are no depth differences between the three salient windows below and the wall, they are not reflected on the depth map, but the three windows above are clearly observed on the depth map. In this case, the depth map will mislead subsequent segmentation. Our BiANet detects multiple objects from RGB images with less noise. The 5th column is also a multi-object scene. The bottom half of the depth map is confused with the interference from the ground. Thanks to the top-down level-by-level refining of uncertain regions, BiANet detects most of the details,

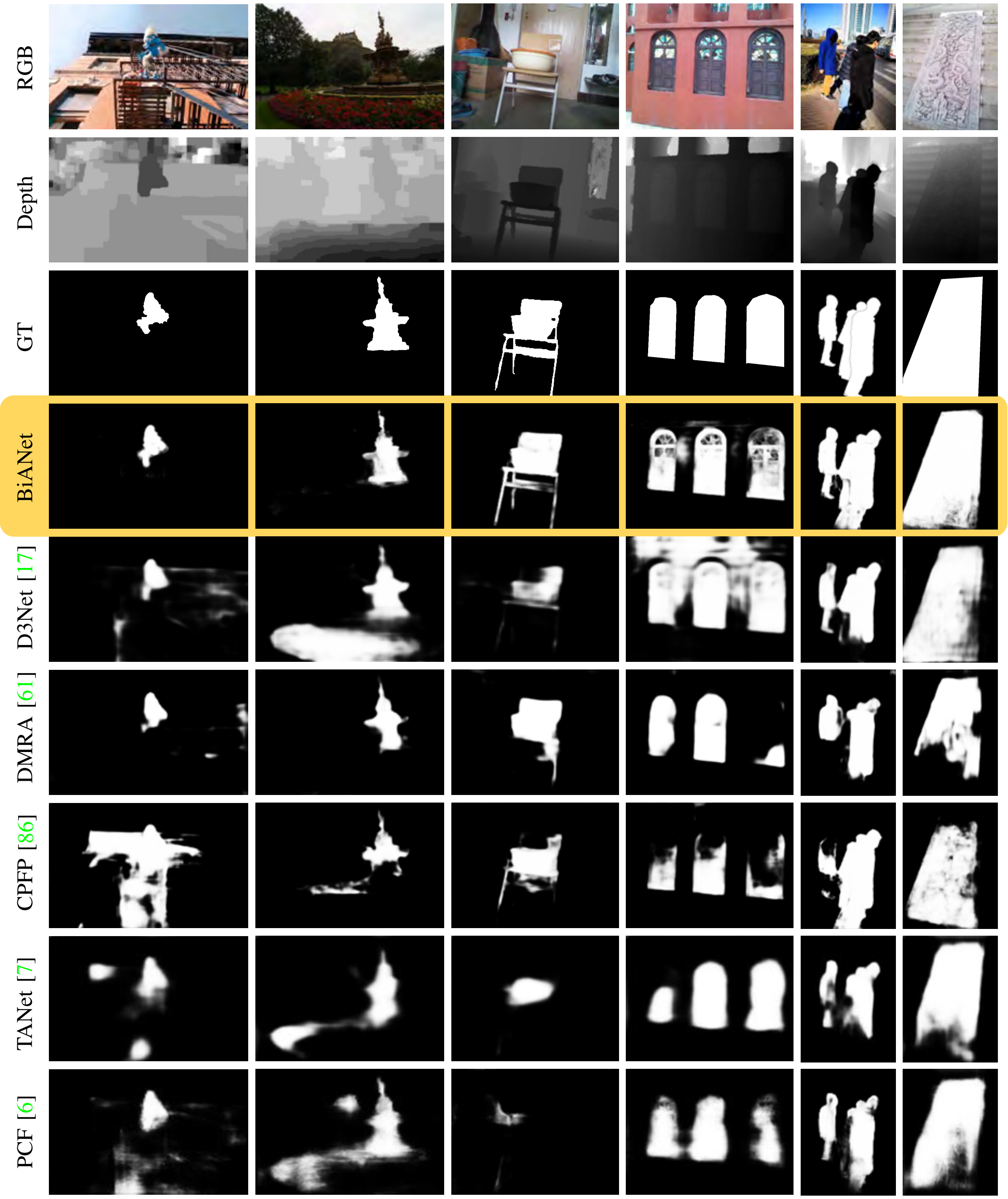

Fig. 6. **Visual comparison of BiANet with other top 5 methods**. The inputs include difficult scenes of tiny objects (column 1), complex background (column 1 and 2), complex texture (column 3), low contrast (column 2 and 6), low-quality or confusing depth (column 2, 4, and 6), and multiple objects (column 4 and 5).

such as the leg area and the person in the middle. The last column is a large-scale object whose color and depth map are not distinguished. Large scale, low color contrast, and lack of discriminative depth information make the scene very challenging. Fortunately, our BiANet is robust in this scene.

C. Ablation Study and Analysis

In this section, we mainly investigate: 1) the benefits of bilateral attention mechanism to our BiANet; 2) the effectiveness of BAM in different levels to our BiANet for RGB-D SOD; 3) the further improvements of MBAM in different levels to our BiANet; 4) the benefits of combining BAM and MBAM for RGB-D SOD; 5) the impact of different backbones to our BiANet for RGB-D SOD; 6) the robustness in detecting non-frontmost objects.

1) Effectiveness of Bilateral Attention: We conduct ablation studies on the *NJU2K* [34] and *STERE* [50] datasets to investigate the contributions of different mechanisms in the proposed

TABLE II

ABLATION ANALYSIS FOR THE PROPOSED ARCHITECTURE ON THE *NJU2K* AND *STERE* DATASETS. THE CANDIDATE MECHANISMS ARE DEEP INFORMATION (DEP), FOREGROUND-FIRST ATTENTION (FF), BACKGROUND-FIRST ATTENTION (BF), AND MULTI-SCALE EXTENSION (ME). ME IS APPLIED ON THE TOP THREE LEVEL FEATURES

#	Candidates				*NJU2K* [34]		*STERE* [50]	
	Dep	FF	BF	ME	$F_\beta \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$
No. 1					0.881	0.885	0.882	0.893
No. 2	✓				0.903	0.904	0.887	0.894
No. 3	✓	✓			0.908	0.908	0.895	0.901
No. 4	✓		✓		0.910	0.908	0.892	0.900
No. 5	✓	✓	✓		0.915	0.913	0.897	0.903
No. 6	✓	✓		✓	0.913	0.911	0.900	0.904
No. 7	✓		✓	✓	0.912	0.911	0.893	0.902
No. 8		✓	✓	✓	0.905	0.903	0.894	0.901
No. 9	✓	✓	✓	✓	**0.920**	**0.915**	**0.898**	**0.904**

method. These two datasets contain large-scaled samples and varied scenes; thus, evaluating on these two datasets can better reflect the performance of different settings. The baseline model used here contains a VGG-16 backbone and a residual refine structure. It takes RGB images as input without depth information. The performance of our basic network without any additional mechanisms is illustrated in Table II No. 1. Based on the network, we gradually add different mechanisms and test various combinations. These candidates are depth information (Dep), foreground-first attention (FF), background-first attention (BF), and multi-scale extension (ME). In Table II No. 3, by applying FF, the performance is improved to some extent. It benefits from the foreground cues being learned effectively by shifting the attention to the foreground objects. We get a similar accuracy when using the BF only, as shown in No. 4. It excels at distinguishing between salient areas and non-salient areas in the background, and can help to find more complete regions of the salient object in the uncertain background; however, too much attention focusing on the background and without a good understanding of the foreground cues, it leads that sometimes background noise is introduced. When we combine FF together with BF to form our BAM and apply it in all side outputs, the performance boosts. We can see that BAM increases S-measure by 0.9% and maximal F-measure by 1.2% compared with No. 2. When we replace the top three levels BAMs with MBAMs, the performance further improved. In order to further verify the importance of mining the foreground and background cues simultaneously, we remove the background-first or foreground-first attention while keeping other components unchanged, and record the results in No. 6 and No. 7. We can see that without the foreground-first or background-first attention, the performance of the proposed model will be reduced. Moreover, the depth information is also critical to bilateral attention. It provides rich foreground and background relationships. We remove the depth information in No. 8, where the prediction accuracy is damaged.

2) Effectiveness of BAM With Different Levels: In order to verify that our BAM module is effective at each feature level, we apply BAM to each side output of the No. 2 model's feature extractor, respectively. That is, in each experiment, BAM is applied to one side output, while the others undergo several convolution groups without being enhanced by foreground-first/background-first attention maps. Specifically, for replacing BAM, the channel-reduced feature undergoes two convolution groups, each of which consists of a convolutional layer with 32 kernels and a ReLU layer. Then, a single convolutional layer follows the two groups to predict the residual. From Table III, we can see that the BAMs in every layer facilitate a universal improvement on detection performance. In addition, we find that BAM applied in the lower levels contributes more to the results. In order to further confirm our observations, we apply BAM in all the five side-output features as the baseline models. Then, we remove one of the five BAMs, and the performances are shown in Table III. We can see that removing BAM from low-level features will cause performance damage.

3) Effectiveness of MBAM in Different Levels: In Table II, compared with No. 5, No. 9 carry out the multi-scale extension on its higher three levels $\{\mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5\}$. This extension effectively improves the performance of the model. In order to better show the gain of MBAM in each level features, similar to the above section, we apply MBAM to each side output of the No. 2 model, respectively. The experimental results are recorded in Table III. Similarly, we also apply MBAM in all the five side-output features as the baseline model. Then, we remove one of the five MBAM to observe performance loss. It can be seen that different levels of MBAM bring different degrees of improvement to the results. Comparing BAM and MBAM, we can see a more interesting phenomenon that the BAM applied in the lower level brings more improvement while the MBAM applied in the higher level is more effective.

4) Cooperation Between BAM and MBAM: The observation above guides us that when using BAM and MBAM in cooperation, we should give priority to multi-scale expansion of higher-level BAM. Therefore, we expand BAM from top to bottom until all BAMs are converted into MBAMs. We record the final detection performance and calculation cost during the gradual expansion in Table IV. We start from the highest level, and gradually increase the number of MBAMs to three. We can see that the effect on the model is a steady improvement, but the computing cost is also increased. At the lower levels, adding MBAM has no obvious effect. This phenomenon is in line with our expectation. Besides, due to the high resolution, the extension of lower-level BAM will increase the calculation cost and reduce the robustness. The selection of the number of MBAM needs to balance the accuracy and speed requirements of the application scenario. In scenarios with higher speed requirements, we recommend not to use MBAM. Our most lightweight model can achieve ~60 *fps* while ensuring significant performance advantages. The parameter size and FLOPs are superior to the SOTA

TABLE III

Effect of accuracy by adding or removing BAM/MBAM in each side outputs. 'None' Denotes the Baseline Model Without Any BAM/MBAM. That Is, 'None' Is the No.2 of Table II. 'w/ L*i*' Means We Add the BAM/MBAM Into the *i*-Th Level Based on the Baseline Model 'None'. 'All' Is the Baseline Model Which Is With BAM/MBAM in All Levels. 'w/o L*i*' Means We Remove the BAM/MBAM From the *i*-Th Level Based on the Baseline Model 'All'

	Metric	w/ L1	w/ L2	w/ L3	w/ L4	w/ L5	None	w/o L1	w/o L2	w/o L3	w/o L4	w/o L5	All
BAM	$S_\alpha \uparrow$	0.908	0.909	0.908	0.906	0.904	0.904	0.911	0.911	0.913	0.912	0.913	0.913
	$F_\beta \uparrow$	0.910	0.911	0.909	0.905	0.904	0.903	0.914	0.914	0.915	0.915	0.915	0.915
	$E_\xi \uparrow$	0.944	0.945	0.943	0.943	0.941	0.942	0.945	0.948	0.947	0.947	0.948	0.948
	$\mathcal{M} \downarrow$	0.043	0.043	0.044	0.044	0.045	0.046	0.041	0.041	0.041	0.042	0.041	0.041
MBAM	$S_\alpha \uparrow$	0.908	0.909	0.910	0.910	0.910	0.904	0.916	0.916	0.914	0.913	0.912	0.916
	$F_\beta \uparrow$	0.909	0.912	0.909	0.911	0.911	0.903	0.920	0.918	0.916	0.914	0.913	0.919
	$E_\xi \uparrow$	0.944	0.945	0.945	0.946	0.947	0.942	0.951	0.947	0.948	0.945	0.946	0.948
	$\mathcal{M} \downarrow$	0.044	0.043	0.042	0.042	0.042	0.046	0.038	0.039	0.039	0.040	0.040	0.039

TABLE IV

Accuracy and calculation cost analysis for MBAM. ×0 ∼ ×5 Means the Number of MBAMs, Which Are Applied From High Levels to Low Levels. *fps* Denotes Frames per Second. Params Means the Size of Parameters. FLOPs = Floating Point Operations. The Accuracy Metrics F_β and $\mathcal{M}$ Are Evaluated on the *NJU2K* Dataset. The Calculation Cost Metrics *fps* and FLOPs Are Tested at 224 × 224 Resolution. Train means the training time. Note That, ×3 Is the Default Setting in Section IV-B

	×0	×1	×2	×3	×4	×5	D3Net [17]	DMRA [61]
$F_\beta \uparrow$	0.914	0.917	0.918	0.920	0.920	0.919	0.887	0.886
$\mathcal{M} \downarrow$	0.041	0.040	0.040	0.039	0.038	0.039	0.051	0.051
fps↑	∼80	∼65	∼55	∼50	∼42	∼34	∼55	∼40
Params ↓	45.0M	46.9M	48.7M	49.6M	50.1M	50.4M	145.9M	59.7M
FLOPs ↓	34.4G	35.0G	36.2G	39.1G	45.2G	58.4G	55.7G	121.0G
Train ↓	0.58h	0.66h	0.81h	1.05h	1.49h	2.29h	-	-

methods D3Net [17] and DMRA [61]. In scenarios where high accuracy is required, we suggest applying less than three MBAMs on higher-level features.

5) Performances Under Different Backbones: We implement the BiANet based on some other widely-used backbones to demonstrate the effectiveness of the proposed bilateral attention mechanism on different feature extractors. Specifically, in addition to VGG-16 [65], we provide the results of BiANet on VGG-11 [65], ResNet-50 [28], and Res2Net-50 [25]. Compared with VGG-16, VGG-11 is a lighter backbone. As shown in Table I, although the accuracy is slightly lower than VGG-16, it still reaches SOTA with a faster speed. BiANet with stronger backbones will bring more remarkable improvements. For example, when we employ ResNet-50 (like D3Net [17]) as backbone, our BiANet brings 1.5% improvement on *NJU2K* [34] in terms of the MAE compared with the D3Net [17]. When armed with Res2Net-50 [25], BiANet achieves 3.8% improvement on *NJU2K* [34] in terms of the maximal F-measure compared with the SOTA methods.

6) Robustness of Detecting Non-Frontmost Objects: In practical applications, our BiANet does not require the salient object to be frontmost in a scene. BiANet jointly explores the saliency cues from the RGB image and depth map. When the depth map brings distance ambiguity, our BiANet still works well in most cases relying on other cues, such as centrality, shape in depth map, and rich cues from the RGB information. Examples in Figure 7 demonstrate the robustness of our BiANet when handling such scenes.

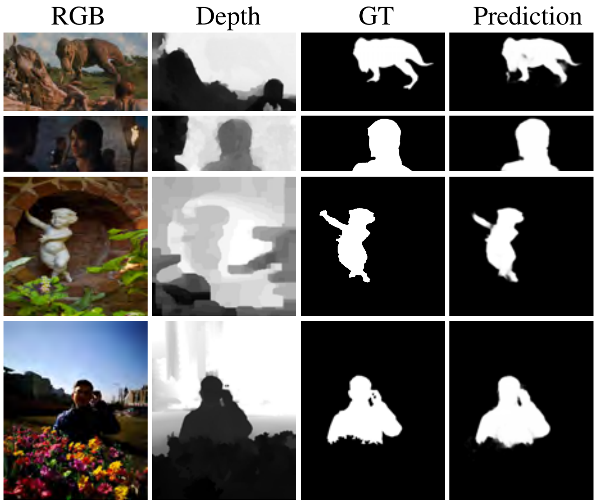

Fig. 7. **Detecting non-frontmost salient objects using BiANet.** The salient objects are behind the rock (see the dinosaur in the first row), not the frontmost person (the second row), or are behind the flowers (the last two rows). We can see that the predictions of BiANet are robust to non-frontmost salient object detection problems.

D. Failure Case Analysis

In Figure 8, we illustrate some failure cases when our BiANet works in some extreme environments. BiANet

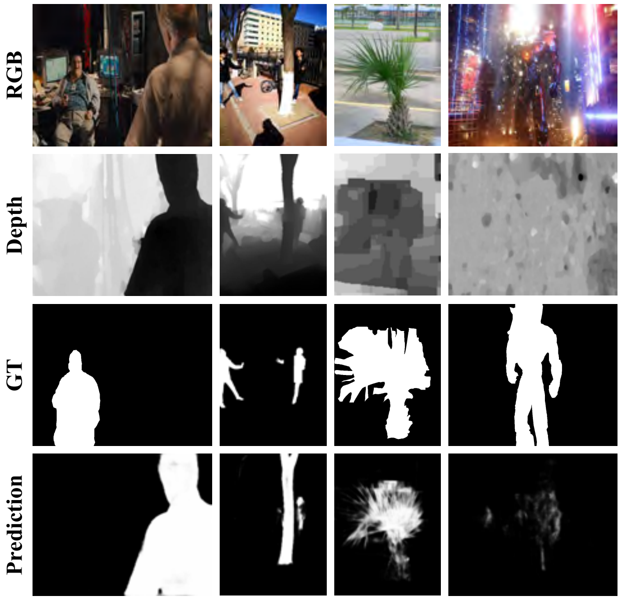

Fig. 8. **Failure cases of BiANet in extreme environments**. In the first two columns, as the objects closer to the observer are not the targets, the depth maps provide misleading information. In the last two columns, the BiANet fails lead by the confusing RGB information and coarse depth maps.

explores the saliency cues bilaterally in the foreground and background regions with the relationship provided by depth information. When the depth map brings distance ambiguity, our BiANet is still robust in most cases depending on other cues, such as centrality, shape in the depth map and rich cues from the RGB information, *etc.*. However, the first two columns in Figure 8 are extreme examples. Specifically, we can see that the target objects are confusing in both the RGB image and depth map.

The other situation that may cause failure is when BiANet encounters coarse depth maps in complex scenarios (see the last two columns). In the third column, the depth map provides inaccurate spatial information, which affects the detection of details. In the last column, the inaccurate depth map and the confusing RGB information make BiANet fail to locate the target object.

V. Conclusion

In this paper, we propose a fast yet effective bilateral attention network (BiANet) for the RGB-D saliency object detection (SOD) task. To better utilize the foreground and background information, we propose a bilateral attention module (BAM) to comprise the dual complementary of foreground-first attention and background-first attention mechanisms. To fully exploit the multi-scale techniques, we extend our BAM module to its multi-scale version (MBAM), capturing better global information. Extensive experiments on six benchmark datasets demonstrated that our BiANet, benefited by our BAM and MBAM modules, outperforms previous state-of-the-art methods on RGB-D SOD in terms of quantitative and qualitative performance. The proposed BiANet runs at real-time speed on a single GPU, making it a potential solution for various real-world applications.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[2] D. Bhowmik and C. Abhayaratne, "Quality scalability aware watermarking for visual content," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5158–5172, Nov. 2016.

[3] D. Bhowmik and C. Abhayaratne, "Embedding distortion analysis in wavelet-domain watermarking," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 4, pp. 1–24, Jan. 2020.

[4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[5] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4296–4307, 2020.

[6] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3060.

[7] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.

[8] 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 Recognit.*, vol. 86, pp. 376–385, Feb. 2019.

[9] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2014, p. 23.

[10] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[11] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and A. Kwong, "Going from RGB to RGBD saliency: A depth-guided transformation model," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3627–3639, Aug. 2020.

[12] 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 Process. Lett.*, vol. 23, no. 6, pp. 819–823, Jun. 2016.

[13] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

[14] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

[15] D.-P. Fan *et al.*, "Re-thinking co-salient object detection," 2020, *arXiv:2007.03380*. [Online]. Available: http://arxiv.org/abs/2007.03380

[16] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, "Taking a deeper look at co-salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2919–2929.

[17] 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 Trans. Neural Netw. Learn. Syst.*, early access, Jun. 3, 2020, doi: 10.1109/TNNLS.2020.2996406.

[18] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8554–8564.

[19] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. ECCV*, 2020, pp. 275–292.

[20] X. Fan, Z. Liu, and G. Sun, "Salient region detection for stereoscopic images," in *Proc. 19th Int. Conf. Digit. Signal Process.*, Aug. 2014, pp. 454–458.

[21] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2343–2350.

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

[23] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3052–3062.

[24] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," 2020, *arXiv:2008.12134*. [Online]. Available: http://arxiv.org/abs/2008.12134

[25] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 30, 2019, doi: 10.1109/TPAMI.2019.2938758.

[26] J. Guo, T. Ren, and J. Bei, "Salient object detection for RGB-D image via saliency evolution," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[27] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 28, no. 6, pp. 2825–2835, Nov. 2018.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

[30] Q. Hou, P.-T. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *Proc. NeurIPS*, 2018, pp. 549–559.

[31] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[32] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[33] P. Jiang, Z. Pan, C. Tu, N. Vasconcelos, B. Chen, and J. Peng, "Super diffusion for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 2903–2917, 2020.

[34] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1115–1119.

[35] C. Jung and C. Kim, "A unified spectral-domain approach for saliency detection and its application to automatic object segmentation," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1272–1283, Mar. 2012.

[36] P. D. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[37] C. Lang, V. T. Nguyen, H. Katti, K. Yadati, S. M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. ECCV*, 2012, pp. 101–115.

[38] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. D. Feng, "Robust saliency detection via regularized random walks ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2710–2717.

[39] G. Li, Y. Gan, H. Wu, N. Xiao, and L. Lin, "Cross-modal attentional context learning for RGB-D object detection," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1591–1601, Apr. 2019.

[40] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, 2020.

[41] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, "GradNet: Gradient-guided network for visual object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6162–6171.

[42] X. Li *et al.*, "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. ECCV*, 2020, pp. 1–24.

[43] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3867–3876.

[44] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, and L. Qing, "Stereoscopic saliency model using contrast and depth-guided-background prior," *Neurocomputing*, vol. 275, pp. 2227–2238, Jan. 2018.

[45] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[46] G. Liu and D. Fan, "A model of visual attention for natural image retrieval," in *Proc. Int. Conf. Inf. Sci. Cloud Comput. Companion*, Dec. 2013, pp. 728–733.

[47] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[48] Y. Liu, J. Han, Q. Zhang, and C. Shan, "Deep salient object detection with contextual information guidance," *IEEE Trans. Image Process.*, vol. 29, pp. 360–374, 2020.

[49] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1007–1013.

[50] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.

[51] J. I. Olszewska, "Active contour based automatic feedback for optical character recognition," in *Proc. BIOSIGNALS*, 2014, pp. 318–324.

[52] J. I. Olszewska, "Active contour based optical character recognition for automated scene understanding," *Neurocomputing*, vol. 161, pp. 65–71, Aug. 2015.

[53] J. I. Olszewska, "Where is my cup?—Fully automatic detection and recognition of textureless objects in real-world images," in *Computer Analysis of Images and Patterns*. Cham, Switzerland: Springer, 2015, pp. 501–512.

[54] J. I. Olszewska, C. De Vleeschouwer, and B. Macq, "Speeded up gradient vector flow B-Spline active contours for robust and real-time tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. 1–905.

[55] J. I. Olszewska, C. De Vleeschouwer, and B. Macq, "Multi-feature vector flow for active contour tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 721–724.

[56] J. I. Olszewska, T. Mathes, C. De Vleeschouwer, J. Piater, and B. Macq, "Non-rigid object tracker based on a robust combination of parametric active contour and point distribution model," in *Proc. VCIP, Int. Soc. Opt. Photon.*, vol. 6508, Jan. 2007, Art. no. 65082A.

[57] J. I. Olszewska and T. L. McCluskey, "Ontology-coupled active contours for dynamic video scene understanding," in *Proc. 15th IEEE Int. Conf. Intell. Eng. Syst.*, Jun. 2011, pp. 369–374.

[58] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019, pp. 8024–8035.

[59] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. ECCV*, 2014, pp. 92–109.

[60] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

[61] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7254–7263.

[62] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[63] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.

[64] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Y. Yang, "Exploiting global priors for RGB-D saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 25–32.

[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[66] H. Song, Z. Liu, H. Du, G. Sun, O. L. Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, Sep. 2017.

[67] G. Te, Y. Liu, W. Hu, H. Shi, and T. Mei, "Edge-aware graph representation learning and reasoning for face parsing," in *Proc. ECCV*, 2020, pp. 258–274.

[68] C.-C. Tsai, W. Li, K.-J. Hsu, X. Qian, and Y.-Y. Lin, "Image co-saliency detection and co-segmentation via progressive joint optimization," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 56–71, Jan. 2019.

[69] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, nos. 1–2, pp. 507–545, Oct. 1995.

[70] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

[71] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1448–1457.

[72] Y. Wang, C. Abhayaratne, R. Weerakkody, and M. Mrak, "Colour space transforms for improved video compression," in *Proc. IWSSIP*, 2014, pp. 219–222.

[73] Y. Wang, X. Zhao, X. Hu, Y. Li, and K. Huang, "Focal boundary guided salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2813–2824, Jun. 2019.

[74] Z. Wang, J. Xu, L. Liu, F. Zhu, and L. Shao, "RANet: Ranking attention network for fast video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3978–3987.

[75] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4142–4150.