Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 8861446, 10 pages
https://doi.org/10.1155/2021/8861446

Research Article

AFI-Net: Attention-Guided Feature Integration Network for RGBD Saliency Detection

Liming Li,[1,2] **Shuguang Zhao**,[1] **Rui Sun**,[2] **Xiaodong Chai**,[2] **Shubin Zheng**,[2] **Xingjie Chen**,[2] **and Zhaomin Lv**[2]

[1] *School of Information Science and Technology, Donghua University, Shanghai 201620, China*
[2] *School of Urban Railway Transportation, Shanghai University of Engineering Science, Shanghai 201620, China*

Correspondence should be addressed to Shuguang Zhao; sgzhao@dhu.edu.cn

Received 28 September 2020; Revised 4 March 2021; Accepted 17 March 2021; Published 31 March 2021

Academic Editor: Anastasios D. Doulamis

This article proposes an innovative RGBD saliency model, that is, attention-guided feature integration network, which can extract and fuse features and perform saliency inference. Specifically, the model first extracts multimodal and level deep features. Then, a series of attention modules are deployed to the multilevel RGB and depth features, yielding enhanced deep features. Next, the enhanced multimodal deep features are hierarchically fused. Lastly, the RGB and depth boundary features, that is, low-level spatial details, are added to the integrated feature to perform saliency inference. The key points of the AFI-Net are the attention-guided feature enhancement and the boundary-aware saliency inference, where the attention module indicates salient objects coarsely, and the boundary information is used to equip the deep feature with more spatial details. Therefore, salient objects are well characterized, that is, well highlighted. The comprehensive experiments on five challenging public RGBD datasets clearly exhibit the superiority and effectiveness of the proposed AFI-Net.

1. Introduction

RGBD salient object detection tries to utilize the pair of RGB and depth images to highlight the visually fascinating regions in RGBD scenarios. Especially with the fast progress of RGBD hardware sensing equipment, such as the traditional Microsoft Kinect, modern smart phones, the advanced time-of-flight camera, and the motion capturing system [1, 2], depth information can be acquired continently and has played an crucial role in many related areas, such as scene understanding [3], semantic segmentation [4], RGBD saliency detection [5], ship detection [6–8], traffic signs detection [9,10], image thumbnails [11], and hand posture detection [12]. Thus, RGBD saliency detection has also received considerable attention, and significant efforts [5, 13–23] have also been exerted on this research area.

However, RGBD saliency models mainly rely on hand-crafted features, such as contrast computation [5, 13], minimum barrier distance computation [22], and the cellular automata model [20]. The performance of these RGBD saliency models degrade largely when handling some complex RGBD scenes with attributions, including small salient objects, heterogeneous objects, cluttered background, and low contrast. This phenomenon can be attributed to the weak representation ability of hand-crafted features in RGBD saliency models. Fortunately, significant progress has been achieved in deep learning theories in the past few years. In particular, convolutional neural networks (CNNs), which provided high level semantic cues, have been applied in RGBD saliency detection successfully [24–35], such as the three stream structure in [34], the fluid pyramid integration in [35], and the complementary fusion in [28].

Though the performance of existing deep learning-based RGBD saliency models is encouraging, they still lose their efficiency when dealing with complex RGBD scenes. Thus, the performance in the area of RGBD saliency detection can still be improved. In addition, some fusion-based RGBD saliency models [5, 14, 15, 19, 27, 28, 32, 33] aim to integrate two modalities, namely, RGB and depth information, through early fusion, middle fusion, and result fusion. These models often result in cross-modal distribution gap or information drop, leading to performance degradation. Meanwhile, the attention mechanism [36] has been widely adopted in many saliency models [37–39], enhancing the saliency detection performance in RGB image scenes. Furthermore, boundary information has been applied in salient object detection [40, 41], providing more spatial details for salient objects.

Thus, this work proposes an innovative end-to-end RGBD saliency model, that is, attention-guided feature integration network (AFI-Net). AFI-Net can extract and fuse features and perform saliency inference. Specifically, our model first extracts multimodal and level deep features, with the pair of RGB and depth images as the input. Then, the attention module, where the attention mechanism [36] is adopted to generate the attention map, enhances the multilevel RGB and depth features, yielding enhanced deep features. Next, these enhanced features (originated from different modalities and levels) are fused hierarchically. Lastly, the RGB and depth boundary features, that is, low-level spatial details, and the integrated feature are combined to perform saliency inference, yielding a high-quality saliency map. Our model focuses on RGBD saliency detection, whereas the existing boundary-aware saliency models [40, 41] focus on performing saliency detection in RGB images.

More importantly, the key advantages of the AFI-Net are the attention model, which indicates the salient objects coarsely, and the boundary information, which provides more spatial details for features. Thus, we can characterize salient objects perfectly in RGBD scenes. The general contributions of AFI-Net are described as follows:

(1) We propose AFI-Net to highlight the salient objects in RGBD images. The AFI-Net has three components, the extraction and fusion of features, and saliency inference.

(2) To sufficiently utilize deep features from different modalities and levels, the attention module is employed to enhance deep features and guide the hierarchical feature fusion. Furthermore, the spatial details are further embedded in the saliency inference step to obtain accurate boundary details.

(3) We perform exhaustive experiments on five public RGBD datasets, and our model achieves the state-of-the-art performance. The experiments also validate the effectiveness of the proposed AFI-Net.

2. Related Works

The pioneer effort [42] defined saliency detection using the center-surround difference mechanism, and succeeding works constructed many saliency models to detect salient objects in natural scene RGB images. Meanwhile, the research on RGBD saliency detection [43, 44] has also been pushed forward significantly in recent decades. Many RGBD saliency models exist, including heuristic models [5, 13–23] and deep learning-based models [24–35, 45], which have achieved encouraging performance. Following, we introduce some of the existing RGBD saliency models.

In [14], color contrast, depth contrast, and spatial bias are combined to generate saliency maps. In [5], luminance-, color-, depth-, and texture-based features are used to produce contrast maps, which are combined to compute for the final saliency map using weighted summation. In [15], the features maps computed by using region grouping, contrast, and location and scale are combined to conduct RGBD salient object detection. In [19], compactness saliency maps computed using color and depth information are aggregated into a saliency map via the weighted summation approach. In [20], color- and depth-based saliency maps are integrated and improved via the linear summation and cellular automata. In [24], various feature vectors, such as local contrast, global contrast, and background prior, are generated and fused to infer the saliency value of each superpixel.

With the wide deployment of CNNs, the performance of RGBD saliency models is significantly advanced. In [25], depth features are combined with appearance features using the fully connected layers, generating high-performance saliency maps. In [27, 28], a two-stream architecture is proposed with a fusion network to detect the complementarities of RGB and depth cues. In [31], two networks, namely, a master network and a subnetwork, are used to obtain deep RGB and depth features, respectively. In [29], RGBD salient object detection is performed using a recurrent CNN. In [35], multilevel features are fused and used to detect salient objects using a fluid pyramid network. In [33], two-stream networks interact to further explore the complementarity of multimodal deep features. In [32], a fusion module is employed to fuse the RGB and depth-based saliency results.

3. Methodology

First, the proposed AFI-Net is introduced in Section 3.1. Then, the feature extraction is presented in Section 3.2. Subsequently, feature fusion and saliency inference are described in Section 3.3. Finally, in Section 3.4, some implementation details are introduced.

3.1. Overall Architecture. Figure 1 summarizes our RGBD saliency model, AFI-Net, which includes a two-stream-based encoder (i.e., feature extraction), a single branch-based decoder (i.e., feature fusion), and saliency inference. Specifically, the entire network is constructed based on VGG-16 [46] with an end-to-end structure. RGB image $\mathbf{I}$ and depth map $\mathbf{D}$ are used as the input to AFI-Net. Here, the initial depth map is encoded into an HHA map $\mathbf{D}$ using [47]. Then, RGB image $\mathbf{I}$ and depth map $\mathbf{D}$ are sent to the two-stream network. Thus, we can obtain multilevel initial deep RGB features $\{\mathbf{AF}_i\}_{i=1}^{5}$ and deep depth features $\{\mathbf{DF}_i\}_{i=1}^{5}$, which correspond to the different convolutional blocks in each branch. Subsequently, a series of attention modules are deployed to enhance the initial deep features, yielding enhanced deep RGB features $\{\mathbf{AFE}_i\}_{i=1}^{5}$ and deep depth features

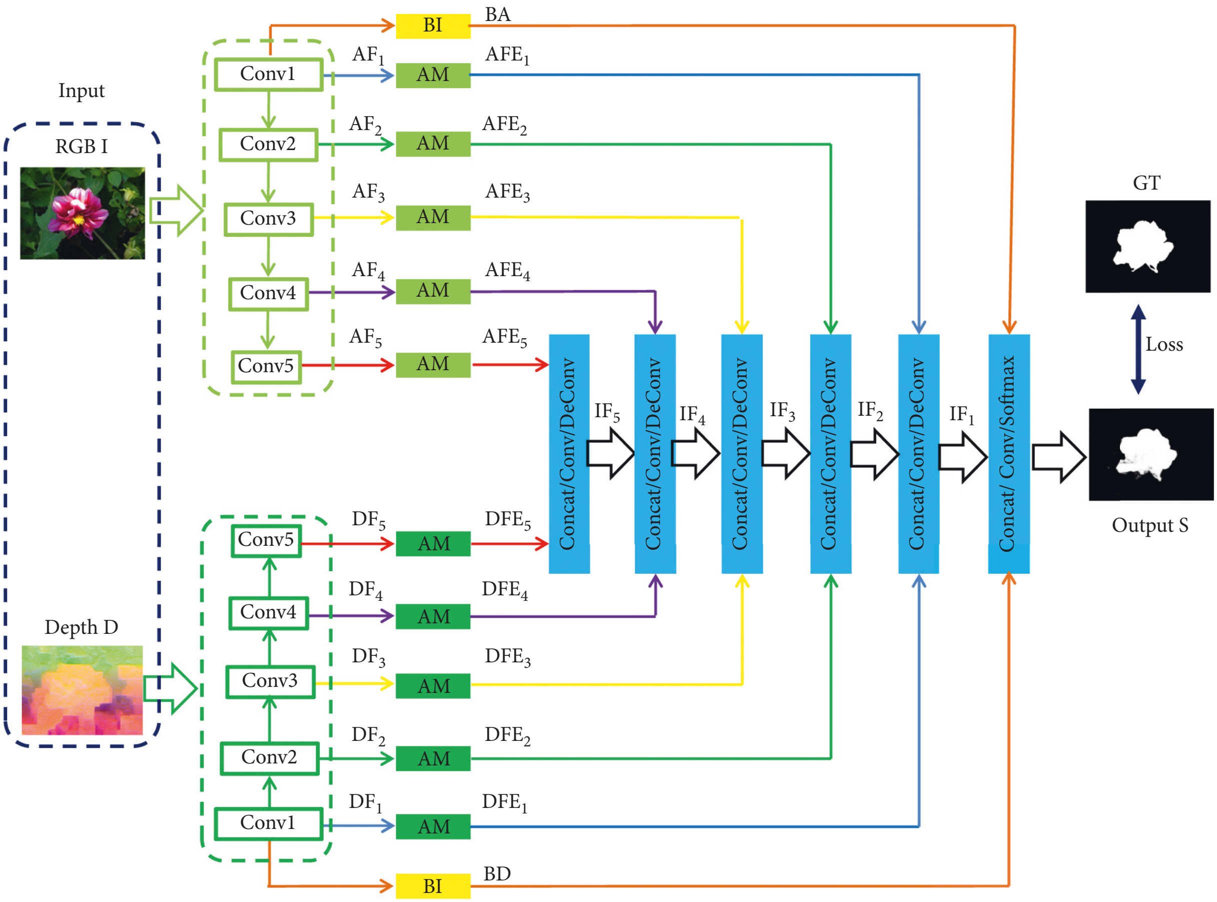


Figure 1: The architecture of the proposed AFI-Net.

$\{\mathbf{DFE}_i\}_{i=1}^5$. Next, the fusion branch is used to integrate the enhanced RGB and depth features hierarchically, and we can obtain integrated deep features $\{\mathbf{IF}_i\}_{i=1}^5$. Finally, the saliency inference module is employed to obtain a saliency map $\mathbf{S}$ by aggregating the boundary information, that is, the low-level spatial details. In Section 3.2, we elaborate the proposed RGBD saliency model, AFI-Net.

3.2. Feature Extraction. The feature extraction branch, namely, an encoder, is a two-stream network containing RGB and depth branches constructed based on VGG-16 [46]. Specifically, the RGB and depth branches have five convolutional blocks with 13 convolutional layers (kernel size = 3×3 and stride size = 1) and 4 max pooling layers (pooling size = 2×2 and stride size = 2). Here, considering the inherent difference of $\mathbf{I}$ and $\mathbf{D}$, the RGB and depth branches have the same structure with different weights. Following this pipeline, we can obtain the initial multiple modalities and the multilevel features including the deep RGB features $\{\mathbf{AF}_i\}_{i=1}^5$ and the deep depth features $\{\mathbf{DF}_i\}_{i=1}^5$, as shown in Figure 1.

On the basis of multimodal features $\{\mathbf{AF}_i\}_{i=1}^5$ and $\{\mathbf{DF}_i\}_{i=1}^5$, we first deploy the attention module, as shown in Figure 2, to further enhance the initial deep features. Formally, we denote each initial deep RGB feature $\mathbf{AF}_i$ or deep depth feature $\mathbf{DF}_i$ as $\mathbf{F}_i$ for convenience. According to Figure 2(b), attention feature $\mathbf{AF}_i$ is formulated as follows:

$$\mathbf{AF}_i = \mathrm{Conv}(\mathbf{F}_i), \tag{1}$$

where Conv denotes a convolutional layer. Then, we can compute the attention weight $(\mathbf{af}_i(w,h))$ at each spatial location using softmax as shown in Figure 2(b):

$$\mathbf{af}_i(w,h) = \frac{e^{\mathbf{AF}_i(w,h)}}{\sum_{(w',h')\in(W,H)} e^{\mathbf{AF}_i(w',h')}}, \tag{2}$$

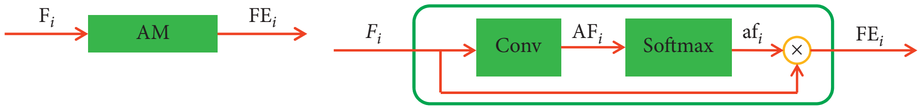
where (w,h) denotes the spatial coordinates of attention feature $\mathbf{AF}_i$ and the width and height of $\mathbf{AF}_i$ are denoted as (W,H). Notably, $\sum_{(w',h')\in(W,H)}\mathbf{af}_i(w',h') = 1$.

After obtaining attention map $\mathbf{af}_i$, initial deep feature $\mathbf{F}_i$ should be selected, which is formulated as follows:

$$\mathbf{FE}_i = \mathbf{DF}_i^{*}\mathbf{af}_i, \tag{3}$$

where $*$ is the Hadamard matrix product in the channel direction and $\mathbf{FE}_i$ is the enhanced deep feature. Thus, we can generate the enhanced deep RGB features $\{\mathbf{AFE}_i\}_{i=1}^5$ and the enhanced deep depth features $\{\mathbf{DFE}_i\}_{i=1}^5$.

3.3. Feature Fusion and Saliency Inference. To integrate the enhanced RGB features $\{\mathbf{AFE}_i\}_{i=1}^5$ and the enhanced depth features $\{\mathbf{DFE}_i\}_{i=1}^5$, the fusion branch, that is, the decoder, is



(a) (b)

Figure 2: Architecture of attention module AM. (a) The thumbnail of AM. (b) The detailed configuration of AM.

deployed to fuse the multimodal and level deep features hierarchically, as shown in Figure 1. Specifically, the hierarchical integration operation is performed as follows:

$$\mathbf{IF}_i = \begin{cases} H\left(\left[\mathbf{DFE}_i, \mathbf{IF}_{i+1}, \mathbf{AFE}_i\right]\right), & i < 5, \\ H\left(\left[\mathbf{DFE}_i, \mathbf{AFE}_i\right]\right), & i = 5, \end{cases} \quad (4)$$

where H denotes the fusion and contains one convolutional layer and one upsampling layer, [.] denotes channel-wise concatenation, and $\mathbf{IF}_i$ is the i^{th} integrated deep feature.

According to the descriptions, we can obtain the first integrated deep feature $(\mathbf{IF}_1)$. On basis of $\mathbf{IF}_i$, we aggregate it with the low-level spatial detail features, that is, the boundary information, to obtain a saliency map. Specifically, as shown in Figure 1, boundary information **BA** and **B D** can be obtained from the bottom layer conv1-2 in the RGB and depth branches, respectively, by using a convolutional layer (1×1), that is, a boundary module (BI box marked in yellow). Subsequently, the saliency prediction is performed by using two convolutional layers (3×3) and one softmax layer. Thus, the saliency inference is written as follows:

$$\mathbf{S} = \mathrm{Fun}\left(\left[\mathbf{IF}_1, \mathbf{BA}, \mathbf{B\ D}\right]\right), \quad (5)$$

where the RGBD saliency map is represented by $\mathbf{S}$, [.] denotes the channel-wise concatenation operation, and Fun refers to the convolutional layers and the one softmax layer.

3.4. Implementation Details. AFI-Net includes feature extraction, feature fusion, and saliency inference. Concretely, $\mathbf{D}_{\text{train}} = \{(\mathbf{I}_n, \mathbf{D}_n, \mathbf{GT}_n)\}_{n=1}^{N}$ is the training dataset, where $\mathbf{I}_n = \left\{\mathbf{I}_n^j,\ j = 1, \ldots, N_p\right\}$, $\mathbf{D}_n = \left\{\mathbf{D}_n^j,\ j = 1, \ldots, N_p\right\}$, and $\mathbf{GT}_n = \left\{\mathbf{GT}_n^j,\ j = 1, \ldots, N_p\right\}$ refer to the RGB image, the depth map, and the ground truth with N_p pixels, respectively. Here, subscript n is dropped, and $\{\mathbf{I}, \mathbf{D}\}$ corresponds to each RGB image and depth map pair. Thus, the total loss can be written as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{W}, \mathbf{b}) = {} & -\beta \sum_{j \in \mathbf{GT}_+} \log P\left(\mathbf{GT}^j = 1 | \mathbf{I}, \mathbf{D}; \mathbf{W}, \mathbf{b}\right) \\ & -(1-\beta) \sum_{j \in \mathbf{GT}_-} \log P\left(\mathbf{GT}^j = 0 | \mathbf{I}, \mathbf{D}; \mathbf{W}, \mathbf{b}\right), \end{aligned} \quad (6)$$

where the kernel weights and bias of the convolutional layers are denoted as $\mathbf{W}$ and $\mathbf{b}$, respectively; $\mathbf{GT}_+$ and $\mathbf{GT}_-$ denote the salient objects and backgrounds, respectively; and β is the ratio of salient objects' pixels in $\mathbf{GT}$, that is, $\beta = |\mathbf{GT}_+|/|\mathbf{GT}_-|$. Furthermore, $P(\mathbf{GT}^j = 1|\mathbf{I}, \mathbf{D}; \mathbf{W}, \mathbf{b})$ is the saliency value of each pixel.

AFI-Net is implemented using the Caffe toolbox [48]. During the training phase, the parameters of the SGD algorithm, such as momentum, base learning rate, minibatch size, and weight decay, are set to 0.9, 10^{-8}, 32, and 0.0001, respectively. Our total iterations are set to 25,000. Furthermore, the learning rate is divided by 10 at the beginning of each 12,500 iterations. The VGG-16 model is used to initialize the weights of the RGB and depth branches. The fusion branch is initialized by using the "msra" method [49]. In addition, the training data used by CPFP [35] is also employed to train our model. The training data contain 1400 pairs of RGB and depth images from NJU2K [16] and 650 pairs of RGB and depth images from NLPR [15]. Obviously, augmentation operations are also adopted, including rotation with angles 90°, 180°, and 270° and mirroring. Finally, the number of training samples is 10250. After the training phase, we can obtain the final model with 131.2 MB. During the test phase, the average processing time per 288 × 288 image is 0.2512 s.

4. Experiments

The public RGBD datasets and the comprehensive evaluation metrics are described in Section 4.1. In Section 4.2, exhaustive quantitative and qualitative comparisons are performed successively. Lastly, the ablation analysis is presented in Section 4.3.

4.1. Experimental Setup. To validate the proposed AFI-Net, we perform comprehensive experiments on five challenging RGBD datasets, namely, NJU2K [16], NLPR [15], STEREO [13], LFSD [50], and DES [14]. NJU2K includes 2003 samples, which are captured from the Internet, daily routines, and so on. From the dataset, 1400 samples are employed for training and 485 samples for testing, that is, "NJU2K-TE." NLPR was constructed by Microsoft Kinect, consisting of 1000 samples, and the salient objects in some samples are more than one. For training the AFI-Net, 650 samples are selected from NLPR to construct the training set, and 300 samples are selected from NLPR to build the testing set, that is, "NLPR-TE." STEREO has 1000 samples, which are used as the testing set. LFSD and DES consist of 100 and 135 samples, which are all used as the testing set. All the datasets are equipped with pixelwise annotation. To compare the RGBD saliency models quantitatively, max F-measure (max F), S-measure (S) [51], mean absolute error (MAE), and max E-measure (max E) [52] are utilized in this paper.

S-measure considers the region aware value (S_r) and the object aware value (S_o) simultaneously, measuring the structural similarity between the ground truth and the saliency map. Referring to [51], the formulation is defined as follows:

Table 1: Quantitative comparisons on five public challenging RGBD datasets.

Metric		CDCP [23]	ACSD [16]	LBE [18]	DCMC [19]	SE [20]	MDSF [21]	DF [24]	AFNet [32]	CTMF [27]	**AFI-net Ours**
NJU2K-TE	$S\uparrow$	0.669	0.699	0.695	0.686	0.664	0.748	0.763	**0.772**	*0.849*	***0.854***
	max $F\uparrow$	0.621	0.711	0.748	0.715	0.748	0.775	**0.804**	0.775	*0.845*	***0.853***
	max $E\uparrow$	0.741	0.803	0.803	0.799	0.813	0.838	**0.864**	0.853	***0.913***	*0.903*
	MAE↓	0.180	0.202	0.153	0.172	0.169	0.157	0.141	**0.100**	*0.085*	***0.073***
NLPR-TE	$S\uparrow$	0.727	0.673	0.762	0.724	0.756	**0.805**	0.802	0.799	*0.860*	***0.878***
	max $F\uparrow$	0.645	0.607	0.745	0.648	0.713	**0.793**	0.778	0.771	*0.825*	***0.864***
	max $E\uparrow$	0.820	0.780	0.855	0.793	0.847	**0.885**	0.880	0.879	*0.929*	***0.933***
	MAE↓	0.112	0.179	0.081	0.117	0.091	0.095	0.085	**0.058**	*0.056*	***0.045***
STEREO	$S\uparrow$	0.713	0.692	0.660	0.731	0.708	0.728	0.757	**0.825**	*0.848*	***0.856***
	max $F\uparrow$	0.664	0.669	0.633	0.740	0.755	0.719	0.757	**0.823**	*0.831*	***0.851***
	max $E\uparrow$	0.786	0.806	0.787	0.819	0.846	0.809	0.847	**0.887**	*0.912*	***0.912***
	MAE↓	0.149	0.200	0.250	0.148	0.143	0.176	0.141	*0.075*	**0.086**	***0.068***
LFSD	$S\uparrow$	0.717	0.727	0.729	0.753	0.692	0.700	**0.791**	0.738	**0.796**	***0.825***
	max $F\uparrow$	0.703	0.763	0.722	**0.817**	0.786	0.783	*0.817*	0.744	0.791	***0.832***
	max $E\uparrow$	0.786	0.829	0.797	0.856	0.832	0.826	**0.865**	0.815	*0.865*	***0.874***
	MAE↓	0.167	0.195	0.214	0.155	0.174	0.190	0.138	**0.133**	*0.119*	***0.097***
DES	$S\uparrow$	0.709	0.728	0.703	0.707	0.741	0.741	0.752	**0.770**	***0.863***	*0.860*
	max $F\uparrow$	0.631	0.756	**0.788**	0.666	0.741	0.746	0.766	0.728	***0.844***	*0.829*
	max $E\uparrow$	0.811	0.850	**0.890**	0.773	0.856	0.851	0.870	0.881	***0.932***	*0.910*
	MAE↓	0.115	0.169	0.208	0.111	0.090	0.122	0.093	**0.068**	*0.055*	***0.050***

The three best results are denoted in bold-italic, italic, and bold fonts.

$$S = \alpha * S_o + (1 - \alpha) * S_r, \tag{7}$$

where α is a balance parameter (here, we set it to 0.5).

F-measure is the weighted harmonic mean of precision and recall and is formulated as follows:

$$F_\beta = \frac{\left(1 + \beta^2\right)\text{precision} \times \text{recall}}{\beta^2\text{precision} + \text{recall}}, \tag{8}$$

where β^2 is set to 0.3. Max *F*-measure could be obtained using different thresholds [0, 255].

E-measure denotes the enhanced-alignment measure considering the local details and the global information. Referring to [52], *E*-measure can be written as follows:

$$\begin{aligned} \xi &= \frac{2\varphi_{\mathrm{GT}}(x, y)^\circ\varphi_{\mathrm{FM}}(\mathrm{x}, \mathrm{y})}{\varphi_{\mathrm{GT}}(x, y)^\circ\varphi_{\mathrm{GT}}(\mathrm{x}, \mathrm{y}) + \varphi_{\mathrm{FM}}(\mathrm{x}, \mathrm{y})^\circ\varphi_{\mathrm{FM}}(\mathrm{x}, \mathrm{y})} \\ E &= \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H} f(\xi), \end{aligned} \tag{9}$$

where $f(\cdot)$ denotes the convex function, $^\circ$ denotes the Hadamard product, and ξ is the alignment matrix.

MAE measures the difference between ground truth **GT** and saliency map **S**:

$$\mathrm{MAE} = \frac{1}{W * H}\sum_{i=1}^{W*H}\left|\mathbf{S}(i) - \mathbf{GT}(i)\right|, \tag{10}$$

where the obtained saliency maps are scaled to [0, 1], W is the width of the saliency map, and H denotes the height.

4.2. Comparison with the State-of-the-Art Models. A comparison is first made on NJU2K-TE, NLPR-TE, STEREO, LFSD, and DES between AFI-Net and nine state-of-the-art RGBD saliency models, namely, CDCP [23], ACSD [16], LBE [18], DCMC [19], SE [20], MDSF [21], DF [24], AFNet [32], and CTMF [27]. The traditional heuristic RGBD saliency models represented by the first six RGBD saliency models and the last three RGBD saliency models are CNNs-based RGBD saliency models. Here, the saliency maps of the other models are provided by the authors or obtained through the source codes. Next, the quantitative and qualitative comparisons are presented. Specifically, Table 1 shows the quantitative comparison results on five RGBD datasets. AFI-Net outperforms the nine state-of-art RGBD saliency models in terms of all the evaluation metrics.

Figure 3 presents the qualitative comparisons on some complex scenes. AFI-Net achieves superior performance over the nine state-of-the-art models. Specifically, the first example presents a box on the ground, where the box in the depth map is indistinctive. The other models shown in Figures 3(e)–3(m) falsely highlight some backgrounds and cannot pop-out the box completely. In the second example, the vase is a heterogeneous object, and its bottom is unclear in the depth map. Our model (Figure 3(d)) performs better than the other models though the top part is not popped-out completely. Like the first example, the third and the fourth examples not only show an unclear depth map but also present a cluttered background. Fortunately, our model still highlights the bird and the cow completely and accurately. The fifth and sixth examples show multiple salient objects. AFI-Net exhibits the best performance, as shown in Figure 3(d). In the seventh example, the man is in the image

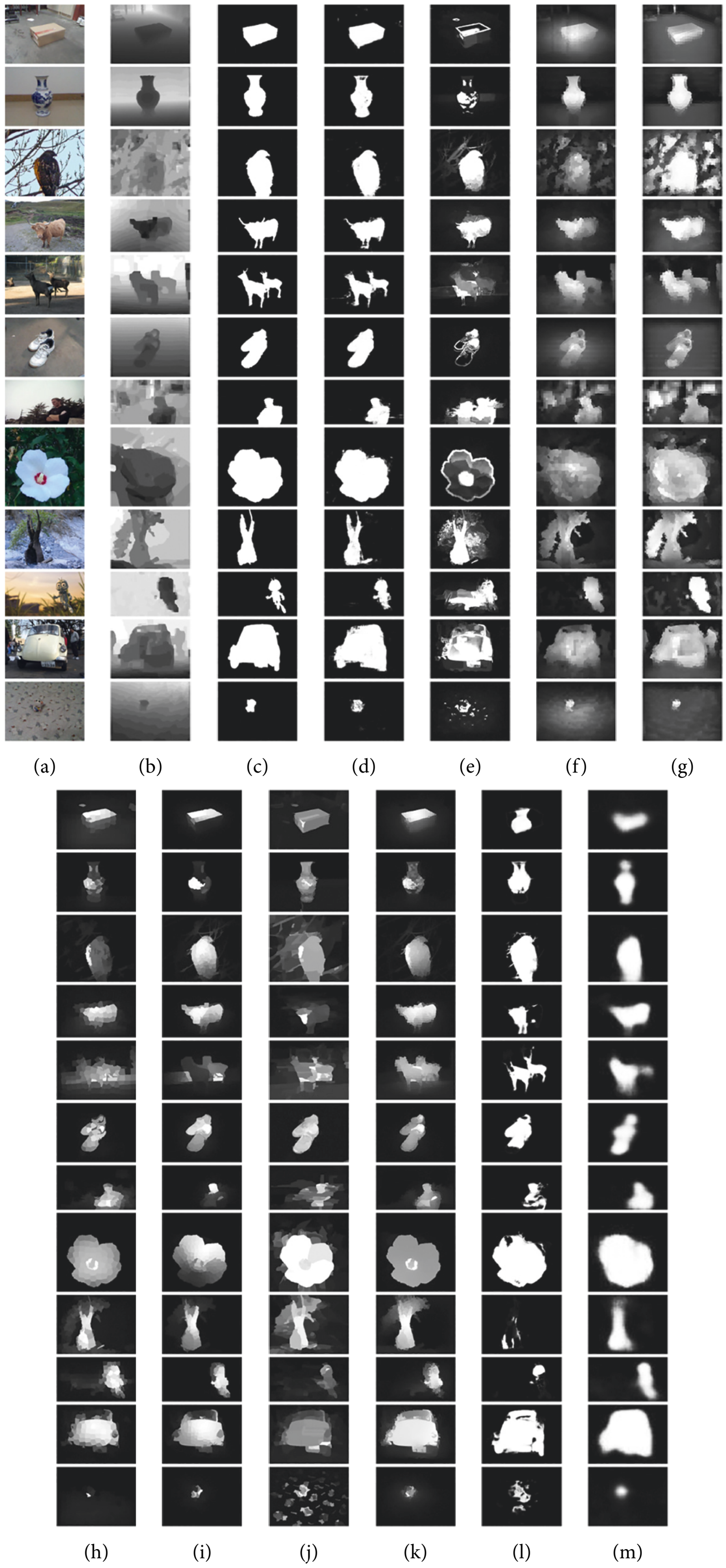


Figure 3: Qualitative comparison results on some challenging scenes. From left to right, (a) RGB, (b) depth, (c) GT, (d) ours, (e) CDCP, (f) ACSD, (g) LBE, (h) DCMC, (i) SE, (j) MDSF, (k) DF, (l) AFNet, and (m) CTMF.

Table 2: Ablation analysis on NJU2K [16] and LFSD [50].

		w/oA	w/oB	AFI-Net
NJU2K-TE	$S\uparrow$	0.824	0.826	**0.854**
	max $F\uparrow$	0.816	0.820	**0.853**
	max $E\uparrow$	0.881	0.884	**0.903**
	MAE$\downarrow$	0.108	0.107	**0.073**
LFSD	$S\uparrow$	0.803	0.798	**0.825**
	max $F\uparrow$	0.803	0.796	**0.832**
	max $E\uparrow$	0.847	0.843	**0.874**
	MAE$\downarrow$	0.131	0.132	**0.097**

The best results are denoted in bold.

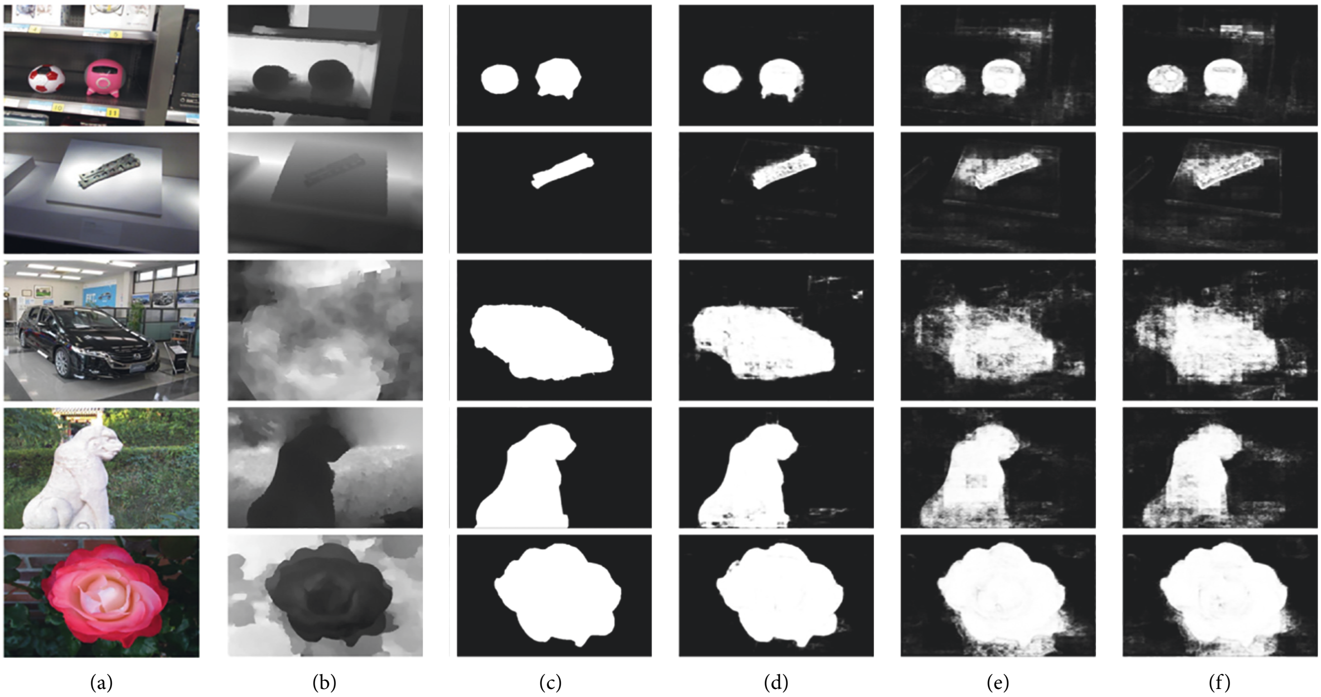


Figure 4: Qualitative comparisons of two variations of AFI-Net on some challenging examples. (a) RGB, (b) depth, (c) GT, (d) AFI-Net, (e) w/oA, and (f) w/oB.

boundary, and its corresponding depth is also unclear. Under this condition, our model still performs better than the others though some backgrounds are also highlighted mistakenly. For the 8th and 11th rows, the salient objects occupy most regions of the images. Our model and the AFNet achieve comparable performance, as shown in Figures 3(d) and 3(m). The 9th and the 10th rows also show a cluttered background. Obviously, AFI-Net still exhibits the best performance as shown in Figure 3(d).

Generally, through the extensive comparison of AFI-Net and nine state-of-the-art models, we can demonstrate the proposed AFI-Net's effectiveness.

4.3. Ablation Studies. Here, the intensive study on some key components in AFI-Net is presented quantitatively and qualitatively. Specifically, the crucial points in AFI-Net include the attention module (AM) and the boundary module (BI), as shown in Figure 1. Therefore, we design two variants of our model, namely, AFI-Net without the attention module (denoted as "w/oA") and the AFI-Net without the boundary module (denoted as "w/oB"). Correspondingly, we perform comprehensive comparisons between our model and the two variants.

First, the quantitative comparison results are presented in Table 2. Clearly, AFI-Net consistently outperforms the two variants, "w/oA" and "w/oB," on two RGBD datasets. Secondly, the qualitative comparison results are presented in Figure 4. AFI-Net (shown in Figure 4(d)) performs better than the two variants (shown in Figures 4(e)) and 4(f)). The results of AFI-Net have well-defined boundaries and highlight the salient objects completely. In contrast, the two variants falsely highlight some backgrounds and cannot detect the salient objects completely.

Overall, the attention and boundary modules play an important role in AFI-Net, enhancing the deep features and equipping them with more spatial details. Meanwhile, the

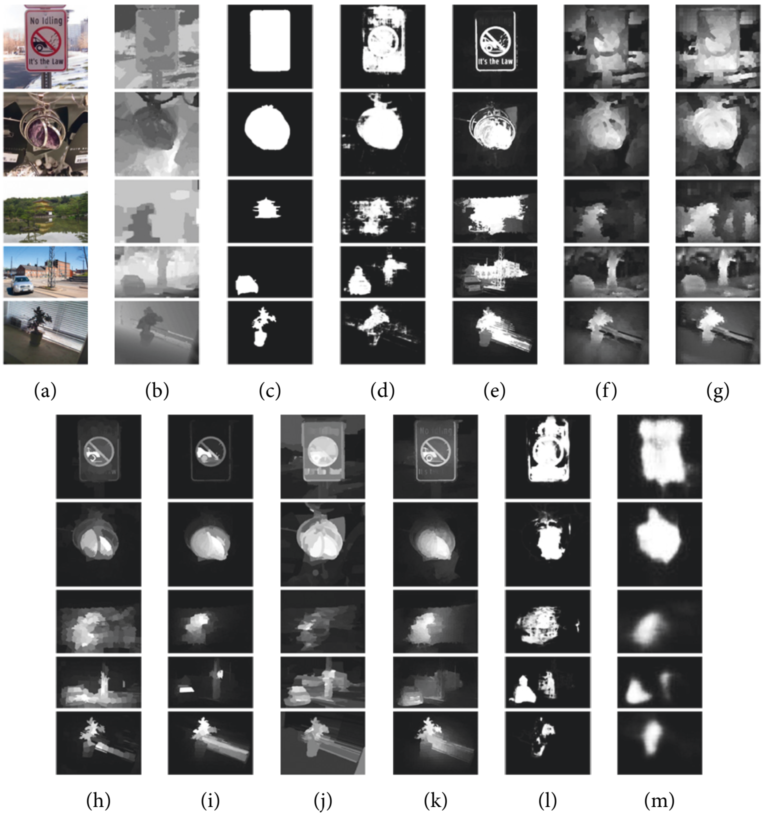


Figure 5: Some typical failure cases. From left to right, (a) RGB, (b) depth, (c) GT, (d) ours, (e) CDCP, (f) ACSD, (g) LBE, (h) DCMC, (i) SE, (j) MDSF, (k) DF, (l) AFNet, and (m) CTMF.

results clearly validate the rationality and effectiveness of the two components in the proposed AFI-Net.

4.4. Failure Case Analysis. In the experiments, we demonstrate the effectiveness and rationality of the proposed AFI-Net. Particularly, Figure 3 shows the qualitative comparison between the proposed AFI-Net and the state-of-the-art saliency models, highlighting the effectiveness of the proposed AFI-Net. However, in some challenging images, our model cannot detect salient objects well, as shown in Figure 5(d). Specifically, in Figure 5, the first example shows a traffic sign, and all models fail to pop-out the salient object. The second example is a pendant, which is highlighted incompletely by most of the models. In the third example, which presents a pavilion, all models falsely pop-out the background regions. In the last two examples, the car and the pot culture cannot be detected accurately and completely. Although our model fails to highlight the salient objects of these examples, it can still pop-out the main part of the salient objects shown in Figure 5(d) better than the other models (shown in Figure 5(e))–5(m)) because our model contains an effective attention module, which covers the main parts of the salient objects. Generally, the research on RGBD saliency detection still faces many difficulties, and the research on the complex scene images is worthy of attention.

5. Conclusion

This work proposes an innovative RGBD saliency model AFI-Net, which can perform feature extraction, feature fusion, and saliency inference. Specifically, the generated initial multimodal and multilevel features are first promoted by a series of attention modules, which select the initial deep features and coarsely indicate the location of salient objects. Then, the hierarchical fusion branch is adopted to fuse the enhanced deep features, which are further combined with low-level spatial detail features (i.e., the boundary information) to perform saliency inference. Thus, the generated saliency maps can highlight salient objects and preserve sharp boundaries. The experiments results on five public RGBD datasets indicate that the proposed AFI-Net obtains superior performance over nine state-of-the-art models.

Data Availability

Previously reported data were used to support this study and are available at https://doi.org/10.1109/cvpr.2012.6247708; https://doi.org/10.1145/2632856.2632866; https://doi.org/10.1007/978-3-319-10578-9_7; https://doi.org/10.1109/icip.2014.7025222; and https://doi.org/10.1109/cvpr.2014.359. These prior studies (and datasets) are cited at relevant places within the text as references [13, 14, 15, 16, 50].

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (nos. 51975347 and 51907117) and the Key Science and Technology Support Project of the Shanghai Science and Technology Commission (no. 18030501300).

References

[1] I. Rallis, I. Georgoulas, N. Doulamis, A. Voulodimos, and P. Terzopoulos, "Extraction of key postures from 3D human motion data for choreography summarization," in *Proceedings of the 2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pp. 94–101, IEEE, Athens, Greece, September 2017.

[2] A. Aristidou, E. Stavrakis, P. Charalambous, Y. Chrysanthou, and S. L. Himona, "Folk dance evaluation using laban movement analysis," *Journal on Computing and Cultural Heritage*, vol. 8, no. 4, pp. 1–19, 2015.

[3] S. Song, S. P. Lichtenberg, and J. Xiao, R.-D. Sun, A rgb-d scene understanding benchmark suite," in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 567–576, IEEE, Boston, MA, USA, June 2015.

[4] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 5199–5208, IEEE, Honolulu, HI, USA, July 2017.

[5] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2625–2636, 2014.

[6] K. Makantasis, E. Protopapadakis, A. Doulamis, and N. Matsatsinis, "Semi-supervised vision-based maritime surveillance system using fused visual attention maps," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 15051–15078, 2016.

[7] T. Cane and J. Ferryman, "Saliency-based detection for maritime object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 18–25, Las Vegas, NV, USA, July 2016.

[8] E. Protopapadakis, K. Makantasis, and N. D. Doulamis, "Maritime targets detection from ground cameras exploiting semi-supervised machine learning," *VISAPP*, no. 2, pp. 583–594, 2015.

[9] W. J. Won, M. Lee, and J. W. Son, "Implementation of road traffic signs detection based on saliency map model," in *Proceedings of the 2008 IEEE Intelligent Vehicles Symposium*, pp. 542–547, IEEE, Eindhoven, The Netherlands, June 2008.

[10] T. Deng, K. Yang, Y. Li, and H. Yan, "Where does the driver look? Top-down-based saliency detection in a traffic driving environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2051–2062, 2016.

[11] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, pp. 2232–2239, IEEE, Kyoto, Japan, October 2009.

[12] Y. Chuang, L. Chen, and G. Chen, "Saliency-guided improvement for hand posture detection and recognition," *Neurocomputing*, vol. 133, pp. 404–415, 2014.

[13] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 454–461, IEEE, Providence, RI, USA, June 2012.

[14] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proceedings of the International Conference on Internet Multimedia Computing and Service,-ICIMCS*, pp. 23–27, ACM, Xiamen, China, July 2014.

[15] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: a benchmark and algorithms," in *Proceedings of the European Conference on Computer Vision, ECCV*, pp. 92–109, Springer, Zurich, Switzerland, September 2014.

[16] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proceedings of the International Conference on Image Processing, ICIP*, pp. 1115–1119, IEEE, Paris, France, October 2014.

[17] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Processing: Image Communication*, vol. 38, pp. 115–126, 2015.

[18] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 2343–2350, IEEE, Las Vegas, NV, USA, July 2016.

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

[20] J. Guo, T. Ren, and J. Bei, "Salient object detection for RGB-D image via saliency evolution," in *Proceedings of the International Conference on Multimedia and Expo, ICME*, pp. 1–6, IEEE, Seattle, WA, USA, July 2016.

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

[22] A. Wang and M. Wang, "RGB-D salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 663–667, 2017.

[23] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Proceedings of the International Conference on Computer Vision, ICCV*, pp. 1509–1515, IEEE, Venice, Italy, October 2017.

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

[25] R. Shigematsu, D. Feng, S. You, and N. Barnes, "Learning RGB-D Salient Object Detection using background enclosure, depth contrast, and top-down features," in *Proceedings of the International Conference on Computer Vision, ICCV*, pp. 2749–2757, IEEE, Venice, Italy, October 2017.

[26] X. Xu, Y. Li, G. Wu, and J. Luo, "Multi-modal deep feature learning for RGB-D object detection," *Pattern Recognition*, vol. 72, pp. 300–313, 2017.

[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 Transactions on Cybernetics*, vol. 48, pp. 3171–3183, 2017.

[28] H. Chen and Y. Li, “Progressively complementarity-aware fusion network for RGB-D salient object detection,” in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 3051–3060, IEEE, Salt Lake City, UT, USA, June 2018.

[29] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, “Salient object detection for RGB-D image by single stream recurrent convolution neural network,” *Neurocomputing*, vol. 363, pp. 46–57, 2019.

[30] H. Chen, Y. Li, and D. Su, “Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection,” *IEEE Transactions on Cybernetics*, vol. 50, no. 11, p. 4808, 2020.

[31] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, “Pdnet: prior-model guided depth-enhanced network for salient object detection,” in *Proceedings of the International Conference on Multimedia and Expo, ICME*, pp. 199–204, IEEE, Shanghai, China, July 2019.

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

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

[34] H. Chen and Y. Li, “Three-stream attention-aware network for RGB-D salient object detection,” *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[35] J. X. Zhao, Y. Cao, D. P. Fan, M. M. Cheng, X. Y. Li, and L. Zhang, “Contrast prior and fluid pyramid integration for RGBD salient object detection,” in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 3927–3936, IEEE, California, CA, USA, June 2019.

[36] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[37] H. Wen, X. Zhou, Y. Sun, J. Zhang, and C. Yan, “Deep fusion based video saliency detection,” *Journal of Visual Communication and Image Representation*, vol. 62, pp. 279–285, 2019.

[38] T. Wang, L. Zhang, S. Wang et al., “Detect globally, refine locally: a novel approach to saliency detection,” in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 3127–3135, IEEE, Salt Lake City, UT, USA, June 2018.

[39] X. Zhang and J. Q. H. L. G. W. Tiantian Wang, “Progressive attention guided recurrent network for salient object detection,” in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 714–722, IEEE, Salt Lake City, UT, USA, June 2018.

[40] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P. M. Jodoin, “Non-local deep features for salient object detection,” in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 6609–6617, IEEE, Honolulu, HI, USA, July 2017.

[41] J. J. Liu, Q. Hou, M. M. Cheng, J. Feng, and J. Jiang, “A simple pooling-based design for real-time salient object detection,” in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 3917–3926, IEEE, California, CA, USA, June 2019.

[42] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[43] D. P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M. M. Cheng, “Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks,” *IEEE Transactions on Neural Networks and Learning Systems*, https://arxiv.org/pdf/1907.06781.pdf, 2020.

[44] T. Zhou, D. P. Fan, M. M. Cheng, J. Shen, and L. Shao, “RGB-D salient object detection: a survey,” *Computational Visual Media*, vol. 7, pp. 1–33, 2020.

[45] W. Zhou, Y. Lv, J. Lei, and L. Yu, “Global and local-contrast guides content-aware fusion for RGB-D saliency prediction,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 175806 pages, 2019.

[46] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *International Conference on Learning Representations*, pp. 1–14, 2014, https://arxiv.org/abs/1409.1556.

[47] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, “Learning rich features from RGB-D images for object detection and segmentation,” in *Proceedings of the European Conference on Computer Vision, ECCV*, pp. 345–360, Springer, Zurich, Switzerland, September 2014.

[48] Y. Jia, E. Shelhamer, J. Donahue et al., “Caffe: convolutional architecture for fast feature embedding,” in *Proceedings of the ACM International Conference on Multimedia, MM*, pp. 675–678, ACM, California, CA, USA, June 2014.

[49] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: surpassing human-level performance on imagenet classification,” in *Proceedings of the International Conference on Computer Vision, ICCV*, pp. 1026–1034, IEEE, Chile, CL, USA, December 2015.

[50] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, “Saliency detection on light field,” in *Proceedings of the Computer Vision and Pattern Recognition, CVPR*, pp. 2806–2813, IEEE, Columbus, OH, USA, June 2014.

[51] D. P. Fan, M. M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: a new way to evaluate foreground maps,” in *Proceedings of the International Conference on Computer Vision, ICCV*, pp. 4548–4557, IEEE, Venice, Italy, October 2017.

[52] D. P. Fan, C. Gong, Y. Cao, B. Ren, M. M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence. IJCAI*, pp. 698–704, Stockholm, Sweden, July 2018.