LIANet: Layer Interactive Attention Network for RGB-D Salient Object Detection

Yibo Han¹, Liejun Wang¹*, Anyu Du¹, Shaochen Jiang¹
¹ College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China;
Corresponding author: Liejun Wang (e-mail: wljxju@xju.edu.cn).

This work was funded by the National Science Foundation of China under grant U1903213, Tianshan Innovation Team of Xinjiang Uygur Autonomous Region under grant 2020D14044.

ABSTRACT RGB-D salient object detection (SOD) usually describes two modes’ classification or regression problem, namely RGB and depth. The existing RGB-D SOD methods use depth hints to increase the detection performance, meanwhile they focus on the quality of little depth maps. In practical application, the interference of various problems in the acquisition process affects the depth map quality, which dramatically reduces the detection effect. In this paper, to minimize interference in depth mapping and emphasize prominent objects in RGB images, we put forward a layered interactive attention network (LIANet). In general, this network consists of three essential parts: feature coding, layered fusion mechanism, and feature decoding. In the feature coding stage, three-dimensional weight is introduced to the features of each layer without adding network parameters, and it is also a lightweight module. The layered fusion mechanism is the most critical part of this paper. RGB and depth maps are used alternately for layered interaction and fusion to enhance RGB feature information and gradually integrate global context information at a single scale. In addition, we also used mixed losses to optimize further and train our model. Finally, a mass of experiments on six standard datasets demonstrated the importance of the method, and a timely detection speed reaches 30 fps on every dataset.

INDEX TERMS RGB-D salient object detection; layered fusion module; parameter-free attention; mixed loss

I. INTRODUCTION

Researchers have proposed many salient target detection models in the past few decades and achieved valuable performance, such as [1-4]. However, most of them use visual cues to process RGB images, which presents insurmountable difficulties in many challenging scenarios. In addition, we live in a natural 3D environment, and the optical system relies heavily on depth information, which can provide enough complementary cues for appearance. Therefore, it is necessary to combine RGB and depth maps to solve the problem of saliency object detection. RGB-D saliency detection has been widely applied, such as image retrieval [5-6], video segmentation [7], pedestrian re-identification [8], visual tracking [9], etc. RGB-D SOD is designed to locate and segment visually significant areas in a scene, and it is typically converted to image-to-mask mapping problems in an end-to-end deep learning pipeline.

In RGB-D SOD, depth maps provide helpful clues such as space structure, 3D layout, target boundaries, etc. To learn effectively, RGB-D SOD usually needs to solve two fundamental problems: 1) how to make full use of abundant depth information for significance prediction, and 2) how to effectively fuse multimodal features between RGB and depth features. This paper focuses on building a non-parametric SOD model that can automatically learn the RGB-D feature interaction structure.

In the last few years, deep learning has developed into a mainstream trend in computer vision applications, particularly pixel-level prediction tasks. The performance of the latest RGB-D SOD methods based on deep learning [10-14] dramatically exceeds that of the traditional RGB-D SOD methods [15-19]. The traditional RGB-D SOD methods generally take the prior information of the depth map and manually extract the distance information to assist significance detection. RGB-D SOD methods based on deep learning are usually data-driven, changing the way people think and adaptively exploring supplemental information in an end-to-end manner.
However, most work often ignores the depth maps quality in traditional approaches or deep learning-based approaches. Therefore, interferences often cause problems for RGB-D SOD in depth maps. For Fig. 1, we demonstrate depth maps of different qualities. It can be seen that an excellent depth map can show distinct boundaries and precise target location information. Poor features cannot provide credible messages and may even negatively influence on the interaction between RGB and depth features, so the detection results of SOD will be affected. In a few parts of the researches, the interference factors in depth maps are taken into account, and the corresponding anti-interference module is proposed. For instance, contrast enhancement networks [20] enhance the contrast; Two-stage depth estimation [21] expands the depth difference of the depth map; depth purification device [22] to judge the depth map quality; The cross-modal attention unit [23] chooses valid areas in the depth map.

From the above method of inspiration, we concentrate on reducing the impact of inaccurate depth maps information and probing effective cross-modal fusions in this article. For this purpose, we propose a new layered interactive fusion module (LIFM) based on Alternating Fusion Unit (AFU). AFU follows an RGB-depth-RGB process for strengthening RGB features, corresponding to modulation and feedback mechanisms. In order to achieve the best efficiency of the AFU module within a single characteristic dimension, we stratified it. At the same time, for the sake of selecting significant features information from multiple AFUs, we perform a channel attention mechanism module for feature re-weighting. In brief, our methods have three outstanding characteristics over [21-23]: flexible alternating fusion unit (AFU), layered structures to grab contexts messages and the feature re-weighting operation.

To sum up, our contributions have the following three aspects:

(1) We propose a new layered interactive fusion module (LIFM), which effectively enhances the cross-modal interaction between RGB and depth features. The RGB-depth-RGB adjustment feedback mechanism in the AFU module successfully eliminates the interference in the depth map and accurately highlights the features of salience objects. And the purpose of the feature re-weighting module is to retain the most valuable information.

(2) We propose an attention module with full three-dimensional weight for feature coding to improve feature representation, and it does not add any parameters in the whole process, and it is a lightweight module.

(3) We propose a mixed loss function to optimize our model further and promote the training of LIANet. A large quantity of experiments has shown that our overall network performs well compared to the other 17 advanced networks.

The following is a brief elaboration of the writing context of this article: The second part introduces the related work. The third part describes the layered interactive fusion module method. The fourth part is the detailed analysis of the experimental performance and results. The fifth part is the summary of the work done in this paper.

II. RELATED WORKS

This section mainly introduces two means related to our proposed method, including RGB-D salient object detection in Section A and relevant models of attentional mechanisms in Section B.

A. RGB-D SALIENT OBJECT DETECTION

Traditional RGB-D saliency target detection methods [24-26] design hand-made features, for instance contrast [26], shape [27], local background closure [24], etc. The difference with the methods described above is that, Li et al. [18] used the distance clue contained in the light field image. Next, Fu et al. [28] and Penge et al. [26] constructed DES and NLPR datasets for RGB-D SOD, respectively. The appearance of these data sets greatly stimulated the study of RGB-D SOD. Gong et al. [29] arrested the depth of prior knowledge. Feng et al. [24] re-weighted the local background closure features with spatial priors and depth information. Guo et al. [30] presented cellular automata to transmit the original saliency graph. Finally, an excellent saliency graph is obtained. Wang et al. [31] acquired the depth significance, depth deviation and 3D prior information from the depth map, and adopted the minimum obstacle distance to optimize the significance map. These traditional methods have reached satisfactory results in some respects, but their generalization is limited by interfering objects and handmade features in depth maps.

Over the years, based on deep learning means have also made significant progress in RGB-D SOD, and many new technical works have been proposed, such as cooperative learning [11], joint learning [32], attention mechanism [33], etc. Li et al. [23] put forward a cross-mode fusion block to go to a step further strengthening RGB features with depth characteristics. Piao et al. [34] came up with a cross-scale module to attain the same goal as above. Peng et al. [26] used a single-flow architecture to directly connect RGB-D pairs as 4-channel inputs to predict significance graphs. Zhao et al. [14] uses single-stream networks and enhanced dual attention for salient object detection. Chen et al. [35] designed a multi-branch network to fuse deep and shallow
cross-modal complementarities in a single path. Zhang et al. [36] chose effective regions from the cross-modal attention module and then they were inset in the network for adaptive training. In this paper, based on our observations, we further utilize depth information, which contains rich geometric prior knowledge. Next, we use depth cues to remove backdrop interference explicitly and propose a practical attentional depth-sensitive attentional module without reference for RGBD salient target detection.

**B. ATTENTIONAL MECHANISM**

Vaswani et al. [36] proposed a self-attention network for natural language modeling. Wang et al. [37] proposed the NL model for learning self-attention in 2D or 3D visual modeling. Nam et al. [38] proposed learning visual and textual attention mechanisms for multimodal reasoning and matching. Mechanical work is also widely used in existing articles combining RGB and depth modes. Piao et al. [34] proposed a new mechanism of circular attention, which can guide cross-modal feature learning and making full use of residual learning and layer-hopping connection. On the other hand, Hu et al. [40] proposed SE to allow a network to capture relevant features and suppress many background activations. Woo et al. [41] proposed BAM and CBAM attention mechanisms combining spatial attention and channel attention in parallel or serial manner. And the computation time of two CBAM is too long. Different from what is described above, we propose a weight attention mechanism module considering the lightweight attributes.

**III. METHOD**

In this part, we first introduce the overall architecture of the proposed RGB-D SOD network based on non-parameter attention in Section A. Then we present the details of a simple no-parameter attention module in Section B. Next, we present the details of the layered interactive fusion module and feature re-weighting in Section C. Finally, the mixed loss functions introduced in Section D. The overall network architecture is shown in Fig. 2.

**A. OVERALL FRAMEWORK OF LIANET**

In our proposed network model, the entire network architecture consists of three parts: encoding, layered interactive fusion and decoding. The coding part consists of an RGB branch, a deep branch and a no-parameter attention module. RGB and depth maps are used alternately for layered interaction and fusion to enhance RGB feature information and gradually integrate global context information at a single scale. See Fig. 2. Considering the...
problem of computational efficiency, we adopted the relatively shallow network VGG16 as the backbone network for encoding feature extraction. We removed the last maximum pooling layer and three fully connected layers. RGB images and depth map obtained characteristic information of different scales by modified dual-stream VGG16. $R_F^{(t)}$ and $D_F^{(t)}$ ($t \in \{1,2,3,4,5\}$) are used to represent the code blocks of extracted RGB maps and depth maps respectively. This structure has been widely used in RGB-D saliency target detection [10-11,42-44]. More importantly, we only use the operation on the feature map of the last convolution layer. In addition, the input sizes for RGB images are $352 \times 352 \times 3$, and the input sizes for depth images are $352 \times 352 \times 1$. The fusion between RGB features and depth features from the coding section is critical. Previous methods have been widely used, such as multiplying [10-11], and noting the connection [44-45]. Just to compare that with the previous approaches, we propose an efficient interaction that can remove the interference features in the depth map. First, we put forward a simple and effective convolutional network model without adding the original network parameters. These features and depth features are then entered into the fusion module (LIFM). In addition, feature re-weighting is implemented to greatly increase the adaptive interaction’s flexibility. Finally, decoding operation and progressive reasoning can gradually restore the resolution of the feature map. Each decoder block consists of two or three convolution layers and one deconvolution layer.

### B. A SIMPLE ATTENTION MODULE

![Diagram](image)

**FIGURE 3.** The detail of the SAM module. We use a 3-D weight to obtain the output feature maps. And the same color represents each channel, spatial position on that feature employing a single scalar.

In general, the attention mechanism module is integrated into each block to refine the output of the previous layers. These methods can generate 1-D or 2-D weights and perform the same treatment on neurons in each channel and spatial location, while this may restrict their ability to learn other discriminations. Therefore, we propose a fully three-dimensional weighted model that is superior to one dimensional or two dimensional to improve these features, as shown in Fig. 3. To achieve attention successfully, we need to estimate the importance of each neuron. In visual processing, active neurons may also inhibit the activity of surrounding neurons, and neurons that exhibit significant spatial inhibitory effects should be given higher priority. As a result of these findings, the energy function is defined as the following:

\[
e_s(w_s, b_s, y, x_i) = (y \odot \hat{s})^2 + \frac{1}{M-1} \sum_{i=1}^{M-1} (y_k - \hat{x}_i)^2, (1)
\]

where, $\hat{s} = w_s + b_s$ and $\hat{x}_i = w_i x_i + b_s$ represent a linear transformation of $s$ and $x_i$. $s$ and $x_i$ represent the target neuron and other neurons of a single channel with input characteristic $X \in R^{CHXW}$. $i$ is the index of spatial dimensions, and $M$ is the number of neurons in that channel. $w_i$ and $b_s$ represent weights and biases respectively. The above formula is equivalent to training the linear separability between neurons $s$ and other neurons in the same channel. For simplicity, we use binary labels and add regular entries. The final function is defined as follows:

\[
e_s^*(w_s, b_s, y, x_i) = \frac{1}{M-1} \sum_{i=1}^{M-1} \left(-1 - (w_s x_i + b_s)\right)^2 + (1 - (w_s x_i + b_s))^2 + \gamma w_s^2, (2)
\]

In the abstract, each channel has $M$ energy functions. Solving all these equations by some iterative solver, such as SGD, is computationally cumbersome. Thus, Eq. (2) can be obtained in the following way:

\[
w_s = -\frac{2(s-u_s)}{(s-u_s)^2+2\bar{\delta}_s^2+2y}, (3)
\]

\[
b_s = -\frac{1}{2} (s + u_s) w_s, (4)
\]

Among them, $u_s = \frac{1}{M-1} \sum_{i=1}^{M-1} x_i$ and $\delta_s^2 = \frac{1}{M-1} \sum_{i=1}^{M-1} (x_i - u_s)^2$ are mean and variance of the surrounding channel neurons. Both Eq. (3) and Eq. (4) are obtained on a single channel, so we can reasonably assume that all pixels in a single channel obey the same distribution. In other words, the mean and variance of all neurons can be calculated and reused for all neurons in this channel [46]. This algorithm avoids the iterative calculation of $u$ and $\delta$ for each position, which greatly reduces the computation amount. Therefore, the minimum energy function is defined as follows:

\[
e_s^* = \frac{4(\bar{\delta}_s^2 + 2y)}{(s - \bar{u})^2 + 2\bar{\delta}_s^2 + 2y}, (5)
\]

where $\bar{u} = \frac{1}{M} \sum_{i=1}^{M} x_i$ and $\bar{\delta}_s^2 = \frac{1}{M} \sum_{i=1}^{M} (x_i - \bar{u})^2$. Eq. (5) indicates that the lower energy, the better the big between neurons $s$ and surrounding neurons, and the better the processing effect of the model. Therefore, the importance of each neuron can be denoted by $1/e_s^*$. Hillyard et al. [47] proposed that attentional regulation in the mammalian brain usually expressed as a gain effect on neuronal reactions. Therefore, we use a scaling operator instead of an additive feature refinement. The following formula can express it:

\[
\hat{x} = \text{sigmoid} \left(\frac{1}{E}\right) \ast X, (6)
\]

where $E$ groups all $e_s^*$ across channel and spatial dimension $s$. Sigmoid is added to limit the value of $E$ to too large.
Knowing the principles of SAM, we show the effects of using the SAM module as shown in Fig. 4. The first line is the original RGB graph, the second line is the truth graph, the third line is the baseline model of our network, and the fourth line is the display result after using SAM. By comparison, the result of line 4 is significantly better than that of line 3. Using the SAM module on the NJUD dataset, the detection range of the class activation graph is larger than that detected in the baseline, that is, the detection results after using SAM contain more active feature information. In the SIP dataset, more noise information is paid attention to by only extracting features from baseline, while the existence of noise information is effectively reduced by SAM, and only significant information is paid attention to. On LFSD and SSD datasets, the edge information of class activation features is clearer. Overall, by adding SAM, the visibility of the class activation feature has been significantly enhanced.

![Image](image.jpg)

**FIGURE 4.** A visual comparison of function activation on the NJUD, LFSD, SIP, and SSD datasets.

### C. LAYERED INTERACTIVE FUSION MODULE

The layered interactive fusion module is the critical part of the network, connecting the encoding network with the decoding network. It filters the interference features and enhances the cross-modal salient features. LIFM is shown in Fig. 5. There are three main components in LIFM: (1) the layered branch; (2) An alternating fusion unit; (3) A feature re-weighting module. Next, we will introduce them in the detail.

1) **LAYERED BRANCH**

Take layer 1 of the backbone network as an example. The size of RGB features is represented as $R_{h_t \times w_t}$, and the same is true of depth signatures. First, four identical hierarchical expansion convolutions are performed for features with expansion rates of 1, 3, 5, and 7, as shown in the following formula:

$$
\begin{align*}
    f_r^{t,j} &= \text{conv}(F_{R_t}^r, W_{R_t}^{t,j}, r_j), \quad j \in \{1,2,3,4\}, \\
    f_d^{t,j} &= \text{conv}(F_{D_t}^d, W_{D_t}^{t,j}, r_j), \quad j \in \{1,2,3,4\},
\end{align*}
$$

where \( \{f_r^{t,j}, f_d^{t,j}\} \in R^{c_t \times h_t \times w_t} \) are layered features of the RGB branch and depth branch, \( W_{R_t}^{t,j} \) represents the number of parameters of different convolution kernel sizes, \( r \) represents expansion rates of various sizes, and \( j \) is branch index.

2) **ALTERNATION FUSION UNIT**

In Fig. 5, there are a few channel attention modules (CAM) \([41]\) and spatial attention (SA) operations in AFU. The following formula is expressed as:

$$
\begin{align*}
    CAM(F) &= \alpha \left( \text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F)) \right), \\
    SA(F) &= \alpha(\text{MaxPool}(F)),
\end{align*}
$$

where \( \alpha \) denotes the sigmoid function, MLP denotes a multilayer perception, AvgPool and MaxPool denote average pooling and max pooling operations respectively. \( F \) is the input feature. In AFU, we first add CAM to the RGB branch, which is used to enhance the characteristics of the RGB branch to get \( AC_{R_t}^{t,j} \). Then we add SA to the enhanced branch features to get \( AS_{R_t}^{t,j} \). Unlike \( AC_{R_t}^{t,j} \), \( AS_{R_t}^{t,j} \) is used to modulate the interference characteristics of the depth branch. We also added a modulation depth branching feature to enhance contrast and realize the RGB-D depth modulation feature. This modulation process is defined as:

$$
\begin{align*}
    f_{rd}^{t,j} &= \text{SA}(\text{CAM}(f_{R_t}^{t,j}) \odot f_{R_t}^{t,j}) \odot f_{d}^{t,j} \oplus f_{d}^{t,j},
\end{align*}
$$

where \( \odot \) is the channel-wise multiplication, \( \odot \) is the element-wise multiplication, and \( \oplus \) is the element-wise summation.

Conversely, we also use CAM and SA operations to get the \( AC_{D_t}^{t,j} \) and \( AS_{D_t}^{t,j} \). \( AS_{D_t}^{t,j} \) is mainly responsible for the RGB branch property, which is called depth RGB feedback mechanism. The process can be expressed as:

$$
\begin{align*}
    f_{dr}^{t,j} &= \text{SA}(\text{CAM}(f_{rd}^{t,j}) \odot f_{rd}^{t,j}) \odot f_{r}^{t,j},
\end{align*}
$$

where \( f_{dr}^{t,j} \) is also the output characteristics of AFU-1.

In AFU, the fusion mode follows the RGB-depth RGB feedback mechanism. This incremental approach more effectively integrates local and global features increasing the interaction between different branches. So, in each branch, \( f_r^{t,j} \) in the Eq. (10) and Eq. (11) can be shown as:
The relationship between the four groups in series. The four groups are combined by element-wise summation. Four features $\mathbf{f}_{\text{AFU}}$, $\mathbf{f}_{\text{CAM}}$, $\mathbf{f}_{\text{SA}}$, and $\mathbf{f}_{\text{AC}}$ are generated from the AFU modules, and then the feature re-weighting is designed to automatically select and focus on different representation of semantic level and spatial position. When these features are fused intuitively, we use element-wise summation to get a global feature $\mathbf{F}_{\text{rd}}$. 

After completing all the feedback mechanisms, we sum up to get a global feature $\mathbf{F}_{\text{rd}}$ the following:

$$F_{\text{rd}} = \text{concat}(f_{\text{AFU}}^{t,1}, f_{\text{AFU}}^{t,2}, f_{\text{AFU}}^{t,3}, f_{\text{AFU}}^{t,4}),$$

where concat(·) shows the concatenation. Through the above series of processing work, $F_{\text{rd}}$ contains abundant salient objects information.

3) FEATURE RE-WEIGHTING

To further optimize the extracted features, we decided to re-weight the $F_{\text{rd}}$. Four features $(f_{\text{AFU}}^{t,1}, f_{\text{AFU}}^{t,2}, f_{\text{AFU}}^{t,3}, f_{\text{AFU}}^{t,4})$ coming out of AFU, their feature maps have the same size and different representation of semantic level and spatial position. When these features are fused intuitively, we use matrix operation generated by an adaptive algorithm to modulate in channel mode, and achieve more valuable feature extraction. As shown in Fig. 6 shows, feature re-weighting is designed to automatically select and focus on more effective information between different groups. Structurally speaking, feature re-weighting is an extended learning phenomenon of CAM [41] in terms of deep supervision and integration. Firstly, there are four groups of outputs are summed from AFU modules, and then the information between the groups is extracted with CAM structure. At the same time, a CAM module is used to extract the relationship between the four groups in series. The feature re-weighting module is expressed as follows:

$$M_{\text{intra}} = \text{CAM}(f_{\text{AFU}}^{t,1}, f_{\text{AFU}}^{t,2}, f_{\text{AFU}}^{t,3}, f_{\text{AFU}}^{t,4}),$$

$$F_m = [f_{\text{AFU}}^{t,1}, f_{\text{AFU}}^{t,2}, f_{\text{AFU}}^{t,3}, f_{\text{AFU}}^{t,4}],$$

$$M_{\text{inter}} = \text{CAM}(F_m),$$

$$F_{\text{out}} = (F_m + \text{repeat}(M_{\text{intra}})) \otimes M_{\text{inter}},$$

where [·] denotes the concatenation of groups of feature maps, the function repeat (·) denotes the operation of repeating the attention map and concatenating in the channel dimension, $\otimes$ denotes element-wise product.

![Feature re-weighting module](image)
D. DETAILS OF THE LOSS FUNCTION

In the field of salient object detection, the clarity of precise edge detection of salient objects is an important challenge. In order to reduce the effect of sample imbalance, we used a mixed loss function during the training phase, composed of the Binary Cross-Entropy (BCE) loss and DICE loss. The total network loss is defined as:

\[ L = L_{bce} + L_{dice}, \]  

(19)

These two loss functions go hand in hand. We use this mixed loss after each decoding module, as shown in Fig. 2. On the one hand, this kind of deep supervision makes our network model converge rapidly, on the other hand, it also improves the precision of significance detection. Each decoder block predicts a significance graph, which is denoted as \( S^{(t)} \). During the training phase, each \( S^{(t)} \) is supervised by a truth graph with BCE and DICE. In other words, Eq. (19) can also be further expressed as:

\[ L = \sum_{t=1}^{5} \left( L_{bce}^{(t)} \left( \text{up}(S^{(t)}), G \right) + L_{dice}^{(t)} \left( \text{up}(S^{(t)}), G \right) \right), \]

(20)

where \( L_{bce}^{(t)}(\cdot, \cdot) \) represents the BCE loss, \( L_{dice}^{(t)}(\cdot, \cdot) \) represents the DICE loss, and \( \text{up}(\cdot) \) represents the bilinear upsampling. BCE loss is widely used in binary classification and segmentation, which is defined as:

\[ l_{bce} = -\sum_{(x,y)} g(x,y) \log(p(x,y)) + (1 - g(x,y)) \log(1 - p(x,y)), \]

(21)

where \( g(x,y) \in [0, 1] \) is the ground-truth label of the pixel \((x, y)\) and \( p(x,y) \in [0, 1] \) is the predicted probability of being the salient object. Considering that BCE Loss independently calculates the loss of each pixel and ignores the global features of the image, we also use DICE Loss. DICE coefficient is a similarity measurement function, which is usually used to measure the similarity of two samples and consider the overall similarity. The formula of DICE Loss is:

\[ l_{dice} = 1 - \frac{2 \times \sum_i p_i g_i}{\sum_i p_i + \sum_i g_i}, \]

(22)

\( p_i \) is predicted value, \( g_i \) is the truth value. We also did the ablation of mixed loss in section 4.4.4.

IV. EXPERIMENT

In this section, we introduce some experimental details in Section 4.1. Then, we introduce datasets and evaluation metrics in Section 4.2 more detail. Next, compare performance analysis with other advanced methods in Section 4.3. In the end, we introduce the study of ablation experiments in three different datasets in Section 4.4.

A. IMPLEMENTATION DETAILS

Our code is done with Pytorch, and the device we use is Tesla V100-PCIE-16GB with 8 graphics cards. For the RGB branch and depth branch, we use a relatively shallow network(vgg-16) to extract features. Following [11, 43], LIANet is trained on a composite training set, including 1400 samples from NJU2K [48] dataset and 650 samples from NLPR [26] dataset. When training, the size of our input pictures is 352*352 and then they also go through flipping, clipping and rotating. On the other hand, the magnitude of the learning rate is 1e-4, weight decay to 0.1 and the total number of parameters to 50909506. We set epoch to 60, batch size to 6, and save the model every five training sessions. The whole framework is trained by using stochastic gradient descent (SGD) in an end-to-end manner. As LIANet is a fully convolutional network, it consumes approximately 30 fps when handing input pictures with a spatial resolution of 352×352.

B. DATASETS AND EVALUATION METRICS

We evaluated our experiments on six commonly used RGB-D datasets against 17 other advanced methods. Let’s start with a brief introduction to seven datasets. STERE [49] contains 1000 pairs of stereoscopic Internet images, most of which have various outdoor scenes and objects. LFSF [18] is a light field SOD dataset. It includes 100 indoor and outdoor images and depth maps. Most images contain simple foreground objects but have complex or similar backgrounds. NLPR [26] consists of 1,000 images of 11 kinds of indoor and outdoor scenes. RGBD135 [28] dataset contains 135 images. But most of their images include relatively simple foreground objects and visual scenes. In addition, depth maps are of good quality. The NJUD [48] dataset has 1,985 stereo images collected from the Internet, 3D movies, and photos were taken. Most images show diverse outdoor scenes and foreground objects. SSD [50] is a small-scale dataset with 80 stereo movie frames. The images contain several movie scenes with persons, animals, buildings and so on. They are used as foreground objects. SIP [22], a recently re-leased dataset containing 929 images, is a high-quality depth map proposed to stand out human detection. Based on recent work, we evaluate the six most widely used evaluation indicators of RGB-D from different aspects.

Precision-Recall (PR) curves: In SOD, the PR curve is based on Precision(P) as the ordinate and Recall(R) as the abscissa to evaluate model performance. We obtain H by binary evolution of the saliency target map S. And S is presented by a module. We use the following formula to figure up P and R values by contrasting H and Ground Truth(G) pixel by line:

\[ Precision = \frac{|H \cap G|}{|H|}, \]

(23)

\[ Recall = \frac{|H \cap G|}{|G|}, \]

(24)

F-measure: Under most circumstances, P and R are unable to comprehensively assess the saliency images which are detected by the model. Thereby, people proposed F-measure: the weighted harmonic meaning of recall and precision with non-negative weight \( \beta \). The formula is as following:

\[ F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2Precision + Recall}, \]

(25)

\( \beta^2 \) is generally taken as 0.3, with the experience of [51,52], that is, the weight value of Precision is increased.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Mean Absolute Error (MAE): MAE is to directly compute the average value between the saliency map of model output and ground-truth map. The following formula is used to calculate:

\[
MAE = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(i, j) - G(i, j)|, \tag{26}
\]

W stands for the width of the picture and H stands for height of the picture, and P and G represent the predicted value and ground truth value respectively.

S-measure: structural measures. Region-aware and Object-aware S-measure methods are used to evaluate non-binary foreground maps, which make the evaluation more reliable. This method uses 5 meta-measures on 5 benchmark datasets to prove that the new measurement method is far more than the existing measurement methods and is highly consistent with human subjective evaluation.

\[
S = \theta S_o + (1 - \theta) S_k, \tag{27}
\]

\(S_o\) and \(S_k\) respectively represent the structural similarity of area perception and object perception. The default value of \(\theta\) is 0.5.

E-measure measures local pixel level error and global image level error.

**C. COMPARISON WITH STATE-OF-ART METHODS**

We compare our method with 17 other state-of-the-art methods on six widely used benchmarks, including DF [53], PCF [54], AFNet [55], CPFP [20], MNCI [33], TANet [44], D3Net [22], JL-DCF [13], DMRA [34], S2MA [14], PGAR [56], ICNet [10], DASNet [57], UC-Net [58], DCF [59], DSA2F [60], HAINet [61]. Some of the above methods are trained with subsets of NJU2K [43] and NLPR [44], while others are trained with subsets of NJU2K [48], NLPR [26] and DUTLF-depth [34]. In this article, we use the former. Saliency maps generated by the above method are run through the source code or provided by the author.

1) **VISUAL COMPARISON**

We make a visual comparison with nine of the latest approaches in Fig. 7. It shows the contrast of the saliency detection graph between the most advanced algorithm and our algorithm. From the Fig. 7, we can clearly see that the best significant effects are achieved by using our proposed model among these algorithms. At the same time, the saliency map is more straightforward and more accurate. For example, the last four lines are images with very complex visual scenes, such as complex objects and cluttered backgrounds. As we can see, this scenario is very challenging for most of the previous approaches. However, the proposed method can successfully locate significant objects. We also show large significant objects (line 1), small significant objects (line 2), and multiple significant objects (line 3). Our model is superior to previous methods, which proves the effectiveness of our proposed model.

**FIGURE 7.** Visual comparison with other 9 advanced methods.
We also performed a visual comparison of feature activation on LFSD and SSD datasets. The foreground objects of LFSD dataset are simple, and most of them are images with single significant information. Therefore, the feature representation of pixels and the perception of regional targets are enhanced after class activation, which is particularly obvious in the three images on the left of Fig. 8. SSD is a small dataset with complex background information (animals, buildings, people, etc.). In the process of extracting the feature information of class activation, there will be some interference, so that the feature attention is not significant in other places. This clutter of background information is reduced with SAM. The three images on the right of Fig. 8.

2) COMPARISON WITH ADVANCED METHODS

Fig. 9 and Tab. 1 respectively indicate the quantitative comparison results of the method proposed and other state-of-the-art methods in this article. We compare our model with 9 other methods, namely CPFP [20], D3Net [22], JL-DCF [13], S2MA [14], PGAR [56], ICNet [10], DASNet [57], UC-Net [58], CDNet [62]. Overall, our proposed method performs best on most aggregates and achieves significant performance improvements. In the recent data sets SIP [22] and LFSD [18], the four evaluation indexes we compared all achieved the best performance. On the NJUD [49] dataset, our method is similar to or worse than the best performance. Through the percentage presentation, our method improves the MaxF index by 2% compared with HAINet [61] on the SIP [22] data set. In Tab. 1, we also compare the FPS sizes of each model.
D. ABLATION STUDIES

In this section, we conduct a series of ablation studies to explore further the contribution of each component in the proposed framework of relative importance and particularity. Our backbone network is trained by VGG16. A non-parameter attention module (SAM) and layered fusion mechanism (LIFM) are proposed in the model, and the loss function aspect is modified to optimize our model better. Tab. 2 shows the role of each module.

1) COMPONENT VALIDITY

Here, B represents the baseline with VGG16 as the backbone network, and then through the original RGB and depth map fusion operation, finally decoded, which can be referred to HAINet [61]. It can be seen from the second line that the SAM model we proposed did not increase the number of any parameters, and the evaluation index was improved to some extent. Lines 3, 4, and 5 show the composition of different components, respectively. The best performance was achieved in the sixth combination. In particular, on the

---

**TABLE 1. Quantitative comparisons on six widely used datasets.** We select four commonly used evaluation indicators for my evaluation, including F-measure (max F, the bigger, the better), S-measure (sm, the bigger, the better), MAE (the smaller, the better), and E-measure (em, the bigger, the better). The best results are marked in red font. The blue font is the second best. Purple is the third best.

| Models       | PPS | NJUD MaxF/MAE/em | RGDID35 MaxF/MAE/em | SIP MaxF/MAE/em | SSD MaxF/MAE/em | LFSD MaxF/MAE/em | STERE MaxF/MAE/em |
|--------------|-----|------------------|---------------------|---------------|----------------|-----------------|------------------|
| Ours         | 28  | 0.91/0.964/0.925 | 0.927/0.892/0.952  | 0.879/0.907/0.927 | 0.893/0.898/0.933 | 0.896/0.899/0.953 | 0.885/0.894/0.968 |
| DNet(TIP2019) | 20  | 0.909/0.902/0.946 | 0.885/0.896/0.902  | 0.834/0.857/0.997 | 0.810/0.825/0.959 | 0.891/0.899/0.946 | 0.870/0.854/0.882 |
| DNet(TNLS2020) | 20  | 0.909/0.902/0.946 | 0.885/0.896/0.902  | 0.834/0.857/0.997 | 0.810/0.825/0.959 | 0.891/0.899/0.946 | 0.870/0.854/0.882 |
| ICNet(TIP2020) | 13  | 0.891/0.894/0.952 | 0.871/0.872/0.952  | 0.834/0.857/0.997 | 0.810/0.825/0.959 | 0.891/0.899/0.946 | 0.870/0.854/0.882 |
| DANet(MMD2020) | 40  | 0.926/0.905/0.924 | 0.901/0.877/0.951  | 0.871/0.875/0.945 | 0.840/0.825/0.959 | 0.894/0.899/0.946 | 0.872/0.854/0.878 |
| UC-Net(CVPR2017) | 17  | 0.893/0.915/0.954 | 0.896/0.866/0.951  | 0.867/0.864/0.938 | 0.867/0.841/0.957 | 0.894/0.899/0.946 | 0.870/0.854/0.878 |
| DCF(CVPR2021) | 20  | 0.924/0.911/0.938 | 0.916/0.923/0.955  | 0.891/0.862/0.971 | 0.891/0.862/0.971 | 0.871/0.854/0.878 | 0.870/0.854/0.878 |
| DSAF(CVPR2021) | 34  | 0.921/0.904/0.920 | 0.914/0.898/0.955  | 0.891/0.862/0.971 | 0.891/0.862/0.971 | 0.871/0.854/0.878 | 0.870/0.854/0.878 |
| HAINet(TIP2019) | 20  | 0.876/0.857/0.959 | 0.919/0.919/0.958  | 0.922/0.900/0.952 | 0.921/0.880/0.952 | 0.875/0.875/0.945 | 0.870/0.854/0.878 |
| Ours         | 28  | 0.91/0.964/0.925 | 0.927/0.892/0.952  | 0.879/0.907/0.927 | 0.893/0.898/0.933 | 0.896/0.899/0.953 | 0.885/0.894/0.968 |

**FIGURE 9.** Quantitative comparisons with 9 state-of-the-art methods over 4 datasets.
TABLE 2. Ablation experiments of different components on NJUD, RGBD135 and SSD datasets. B is the baseline. SAM and LIFM are Simple Attention Module and Layered Interactive Fusion Module respectively. Loss stands for total mixed loss.

| # | Settings                  | NJUD         | RGBD135       | SSD           |
|---|---------------------------|--------------|---------------|---------------|
|   |                           | MaxF/sm/MAE  | MaxF/sm/MAE   | MaxF/sm/MAE   |
| 1 | B                         | .876/.857/0.059 | .927/.914/0.024 | .752/.756/1.00 |
| 2 | B + SAM                   | .898/.880/0.050 | .931/.906/0.025 | .754/.759/0.98 |
| 3 | B + SAM + Loss            | .901/.896/0.048 | .932/.904/0.023 | .756/.760/0.96 |
| 4 | B + LIFM                  | .910/.902/0.050 | .926/.899/0.025 | .744/.759/0.93 |
| 5 | B + SAM + LIFM + Loss     | .911/.904/0.042 | .938/.929/0.021 | .759/.771/0.92 |

NJUD [49] dataset, MaxF increased by 3.5% from baseline and MAE decreased by 1.7%. On the whole, each component of LIFMNet provides significant performance gains.

2) IMPACT OF THE NUMBER OF LIFM BRANCHES
To verify the rationality of the hierarchy in LIFM, we changed the number of branches to 1, 2, 4, 6. As shown in Tab. 3, we analyzed the ablation experiment on STERE, SIP and SSD datasets. By comparing the results of branches 1, 2, and 4, we found that the evaluation indicators improved with the increase of the number of branches (MAE: 0.046(1) - 0.042(2) - 0.037(4) on STERE, 0.053(1) - 0.051(2) - 0.049(4) on SIP, 0.098(1) - 0.097(2) - 0.093(4) on SSD). Therefore, we can observe that the more branches there are, the more information will be exchanged and the better detection performance will be. But this performance improvement does not mean that more branches are better, and performance drops when the number of branches is 6. This is because the number of channels in each characteristic size is constant. An increase in the number of branches leads to a decrease in information, which degrades performance. Therefore, we adopt a branch number of 4 model architecture and achieve good results.

TABLE 3. To analyze the rationality of LIFM structure, we provide the structure of 1, 2, 4 and 6 branches, and the bold in each column is the best result.

| Branch | STERE | SIP | SSD |
|--------|-------|-----|-----|
|        | MaxF/sm/MAE | MaxF/sm/MAE | MaxF/sm/MAE |
| 1      | .849/.897/0.046 | .890/.878/0.053 | .742/.748/0.098 |
| 2      | .861/.904/0.042 | .901/.880/0.051 | .747/.752/0.097 |
| 4      | .915/.908/0.037 | .912/.884/0.049 | .759/.771/0.093 |
| 6      | .910/.902/0.050 | .909/.879/0.051 | .744/.759/0.094 |

3) THE ROLE OF FEATURE RE-WEIGHTING MODULE
To investigate the indispensable of feature reweighting modules in LIFM, we also tested the effect of removing feature re-weighting modules, as shown in Tab. 4. For example, the MaxF evaluation index increased by about three points in all three datasets, proving that our feature reweighting operation is essential.

TABLE 4. Ablation analysis of feature re-weighting module. Re-W in the table represents feature re-weighting. “-“ means there is no operation, “+” means there is operation.

| Models | STERE | RGBD135 | SSD |
|--------|-------|---------|-----|
|        | MaxF/sm/MAE | MaxF/sm/MAE | MaxF/sm/MAE |
| - Re-W | .872/.887/0.042 | .889/.801/0.026 | .756/.752/0.094 |
| + Re-W | .907/.904/0.040 | .935/.819/0.022 | .773/.781/0.092 |

4) ABLATION ANALYSIS OF LOSS FUNCTION
In order to study the usefulness of the loss function, we also used three loss functions for comparison: BCE loss, IOU loss, and DICE loss. Different parameter coefficients are also provided to compare mixing loss, as shown in Tab. 5. By comparing the first three lines, it is proved that the mixing effect of two loss functions is better than that of a single BCE loss function, and the total loss of BCE and DICE is better than that of BCE and IOU. Therefore, the total loss function is defined as Eq. (19) in this paper. In lines 4, 5 and 6, the influence results of different coefficients on the loss function are also analyzed. The two loss functions are given different coefficients respectively, and the performance reaches the best when each coefficient is 0.5. MAE decreased by 0.8% in both LFSD and SSD datasets. In particular, MaxF improved 4.7% on SSD datasets. In general, our loss function is very beneficial.

V. CONCLUSIONS
In this paper, we propose an effective layered fusion network for the application of RGB-D saliency target detection. LIANet is designed based on the coder and decoder structure. A simple and effective attention module is added to the coding architecture to significantly extract edge information more without increasing network parameters. The most important part of this paper is the layered interaction fusion module, which is responsible for enhancing the interaction between cross-modal features, that is, filtering out the interference factors in the depth features and strengthening the prominent objects of RGB features. Finally, the significance graph is obtained by the decoding module. Though a comprehensive
analysis of comparative experiments and ablation studies, LIANet is competitive with other advanced methods and strikes a balance between accuracy and speed.

REFERENCES

[1] N. Liu and J. Han, “Dissnet: Deep hierarchical saliency network for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 678–686.

[2] N. Liu, J. Han, and M.-H. Yang, “Picnet: Learning pixel-wise contextual attention for saliency detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 3089–3098.

[3] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and Cheng, “Egnet: Edge guidance network for salient object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8778–8787.

[4] X. Qin, Z. Zhang, C. Huang, and C. Gao, “Basnet: Boundary-aware salient object detection;” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7479-7489.

[5] G. Yue, M. Wang, D. Tao, R. Ji, and Q. Dai, “3d object retrieval and recognition with hypergraph analysis.” IEEE Trans. Image Process., vol. 21, pp. 4290–4303, 2012.

[6] G. Liu, and D.-P. Fan, “A model of visual attention for natural image retrieval,” in International Conference on Information Science and Cloud Computing Companion, pp. 728-733, 2013.

[7] D. Fan, W. Wang, M. Cheng, and J. Shen, “Shifting more attention to video salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8554-8564.

[8] R. Zhao, and X. Wang, “Unsupervised salience learning for person re-identification;” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 3586-3593.

[9] S. Hong, T. You, S. Kwalk, and B. Han, “Online tracking by learning discriminative saliency map with convolutional neural network,” in Proceedings of the 32nd International Conference on Machine Learning, pp. 597-606, 2015.

[10] G. Li, Z. Liu, Y. Wang, and H. Ling, “ICNet: Information conversion network for RGB-D based salient object detection,” IEEE Trans. Image Process., vol. 29, pp. 4873-4884, 2020.

[11] G. Li, Z. Liu, Y. Wang, and H. Ling, “Cross-modal weighting network for RGB-D salient object detection,” in Proceedings of the 2020 ECCV Computer Vision, Aug. pp. 665-681, 2020.

[12] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, “A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection,” in IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9057–9066.

[13] 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,” IEEE Trans. Pattern Anal. 2020.

[14] N. Liu, N. Zhang, and J. Han, “Learning selective self-mutual attention for RGB-D saliency detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 13753–13762.

[15] Y. Fang, J. Wang, M. Narwaria, P. Callet, and W. Lin, “Saliency detection for stereoscopic images,” IEEE Trans. Image Process., vol. 23, pp. 2625-2636, 2014.

[16] X. Fan, Z. Liu, and G. Sun, “Saliency region detection for stereoscopic images,” in Proc. 19th Int. Conf. Digit. Signal Process., Aug. 2014, pp. 454–458.

[17] H. Song, Z. Liu, H. Du, G. Sun, and C. Bai, “Saliency detection for RGBD images,” in Proceedings of the 7th International Conference on Internet Multimedia Computing and Service, Aug. 2015; pp. 72:1-72:4.

[18] N. Li, J. Ye, Y. Ji, H.; Ling, and J. Yu, “Saliency detection on light field,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2014, pp. 2806–2813.

[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 Process. Lett. vol. 23, pp. 819-823, 2016.

[20] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, and L. Zhang, “Contrast prior and fluid pyramid integration for RGBD salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3927-3936.

[21] 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.

[22] D.-P. Fan, Z. Lin, Z. Zhang, and M. Zhu, “Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks,” IEEE Trans. Neural Networks Learn. vol. 32, pp. 2075-2089, 2021.
[23] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, “Select, supplement and focus for RGB-D saliency detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3469–3478.
[24] 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.
[25] C. Lang, T.-V. Nguyen, H. Katti, M. Kankanhalli, and S. Yan, “Depth matters: Influence of depth cues on visual saliency,” in Proceedings of the 2012 ECCV Computer Vision, Florence, Italy, Oct. pp. 101–115, 2012.
[26] H. Peng, W. Xiong, W. Hu, and R. Ji, “Rgb-d salient object detection: a benchmark and algorithms,” in Proceedings of the 2014 ECCV Computer Vision, Zurich, Switzerland, Sept. pp. 92-109, 2014.
[27] A. Ciptadi, T. Hermans, J.-M. Rehg, “An indepth view of saliency,” In Brit. Mach. Vis. Conf.Georgia Institute of Technology, 2013.
[28] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, “Depth enhanced saliency detection method,” in International Conference on Internet Multimedia Computing and Service, Xiamen, China, July 10-12, 2014.
[29] J. Ren, X. Gong, L. Yu, and W. Zhou, “Exploiting global priors for RGB-D saliency detection,” in IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 25–32.
[30] J. Guo, T. Ren, J. Bei, “Saliency object detection for RGB-D image via saliency evolution,” in International Conference on Multi-media and Expo, June 11-15, 2016; pp. 1–6.
[31] A. Wang, and M. Wang, “RGB-D salient object detection via minimum barrier distance transform and saliency fusion,” IEEE Signal Process. Lett. vol. 24, pp. 663-667, 2017.
[32] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, “A single stream network for robust and real-time rgb-d salient object detection,” in Proceedings of the 2020 ECCV Computer Vision, Aug. 2020; pp. 646-662.
[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 Recognit. vol. 86, pp. 376-385, 2019.
[34] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, “Depth-induced multi-scale recurrent attention network for saliency detection,” in Proceedings of the 2019 IEEE Computer Vision, October 27–November 2, 2019; pp. 7253-7262.
[35] C. Li, R. Cong, S. Kwong, H. Fu, and G. Zhu, “ASIF-net: Attention steered interweave fusion network for RGB-D salient object detection,” IEEE Trans. Cybern. vol. 51, pp. 88-100, 2021.
[36] A. Vaswani, N. Shazeer, L. Jones, and L. Kaiser, “Attention is all you need,” Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, pp. 5998-6008.
[37] X. Wang, R.-B. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 7794–7803.
[38] H. Nam, J.-W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2017, pp. 2156–2164.
[39] D. Liu, K. Zhang, and Z. Chen, “Attentive cross-modal fusion network for RGB-D saliency detection,” IEEE Trans. Multim. vol. 23, pp. 967-981, 2021.
[40] J. Hu, L. Shen, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” IEEE Trans. Pattern Anal. vol. 42, pp. 2011-2023, 2020.
[41] S. Woo, J. Park, J.-Y. Lee, and I.-S. Kweon, “Cbam: Convolutional block attention module,” in Proceedings of the 2018 ECCV Computer Vision, Sept. 2018, pp. 3-19.
[59] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, and S. Yao, “Calibrated RGB-D Salient Object Detection,” in IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 9471–9481.

[60] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, “Deep RGB-D Saliency Detection with Depth-Sensitive Attention and Automatic Multi-Modal Fusion,” in IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 1407–1417.

[61] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, “Hierarchical Alternate Interaction Network for RGB-D Salient Object Detection,” IEEE Trans. Image Process. vol. 30, pp. 3528-3542, 2021.

[62] W. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, “CDNet: Complementary Depth Network for RGB-D Salient Object Detection,” IEEE Trans. Image Process. vol. 30, pp. 3376-3390, 2021.