International Journal of Computational Science and Engineering

ISSN online: 1742-7193 - ISSN print: 1742-7185
https://www.inderscience.com/ijcse

JALNet: joint attention learning network for RGB-D salient object detection

Xiuju Gao, Jianhua Cui, Jin Meng, Huaizhong Shi, Songsong Duan, Chenxing Xia

DOI: 10.1504/IJCSE.2024.10061860

Article History:

Received:	19 May 2022
Last revised:	17 July 2022
Accepted:	20 July 2022
Published online:	25 January 2024

JALNet: joint attention learning network for RGB-D salient object detection

Xiuju Gao

School of Electrical and Information Engineering,
Anhui University of Science and Technology,
Huainan, Anhui, China
Email: xjgao@aust.edu.cn

Jianhua Cui*, Jin Meng and Huaizhong Shi

Anyang Cigarette Factory,
China Tobacco Henan Industrial Co., Ltd.,
Anyang, Henan, China
Email: ay_zscjh@163.com
Email: ay_mengj@126.com
Email: ay_zsshz@126.com
*Corresponding author

Songsong Duan and Chenxing Xia

College of Computer Science and Engineering,
Anhui University of Science and Technology,
Huainan, Anhui, China
Email: d15180632812@163.com
Email: cxxia@aust.edu.cn

Abstract: The existing RGB-D saliency object detection (SOD) methods mostly explore the complementary information between depth features and RGB features. However, these methods ignore the bi-directional complementarity between RGB and depth features. From this view, we propose a joint attention learning network (JALNet) to learn the cross-modal mutual complementary effect between the RGB images and depth maps. Specifically, two joint attention learning networks are designed, namely, a cross-modal joint attention fusion module (JAFM) and a joint attention enhance module (JAEM), respectively. The JAFM learns cross-modal complementary information from the RGB and depth features, which can strengthen the interaction of information and complementarity of useful information. At the same time, we utilise the JAEM to enlarge receptive field information to highlight salient objects. We conducted comprehensive experiments on four public datasets, which proved that the performance of our proposed JALNet outperforms 16 state-of-the-art (SOTA) RGB-D SOD methods.

Keywords: salient object detection; depth map; bi-directional complementarity.

Reference to this paper should be made as follows: Gao, X., Cui, J., Meng, J., Shi, H., Duan, S. and Xia, C. (2024) 'JALNet: joint attention learning network for RGB-D salient object detection', *Int. J. Computational Science and Engineering*, Vol. 27, No. 1, pp.36–47.

Biographical notes: Xiuju Gao is currently an assistant at the College of Electrical and Information Engineering, Anhui University of Science and Technology. Her research interests include image/video processing and computer vision.

Jianhua Cui is currently a technician at the Anyang Cigarette Factory, China Tobacco Henan Industrial Co., Ltd. and his job title is an engineer. His research interests include deep learning, edge computing, visual recognition and web services.

Jin Meng is currently a technician at the Anyang Cigarette Factory, China Tobacco Henan Industrial Co., Ltd. and her job title is an engineer.Her research interests include computer vision, image processing and machine learning, etc.

Huaizhong Shi is currently an electrical maintenance team leader at the Anyang Cigarette Factory, China Tobacco Henan Industrial Co., Ltd. and his job title is an engineer. His research interests include deep learning, data mining and image processing, etc.

Songsong Duan is currently pursuing his MS at the Anhui University of Science and Technology. His current research interests include computer vision, and salient detection.

Chenxing Xia is currently an Associate Professor at the College of Computer Science and Engineering, Anhui University of Science and Technology. His research interests include saliency detection, computer vision, and depth prediction.

1 Introduction

Salient object detection (SOD) simulates human visual attention mechanisms to locate the most significant and notable objects or regions in a scene. As one of the most basic and important pixel-level density prediction tasks in computer vision, it has been widely used in many down-stream tasks, e.g., image retrieval (Gao et al., 2012), visual tracking (Hong et al., 2015), medical image segment (Fan et al., 2020c), and person re-identification (Martinel et al., 2015). Most SOD methods (Hou et al., 2017; Qin et al., 2019b; Liu et al., 2021; Xia et al., 2020) detect the salient objects based on a single RGB image. However, it is hard for these methods to completely highlight the salient objects and preserve their rich edge details in challenging and complex scenarios, e.g., similar appearance and texture between the foreground and the background [as shown in Figure 1(a)], transparent objects [as shown in Figure 1(c)] and cluttered background [as shown in Figure 1(b)]. Recently, depth information has received the attention of many researchers due to the accessibility of depth information by some devices, e.g., Kinect, iPhone XR, Huawei Mate30, which can provide many beneficial and complementary cues to the RGB images, such as spatial structure and boundary information. Hence, more and more researchers introduce depth information into the field of SOD as additional complementary information for the RGB images to improve the performance of SOD, named RGB-D SOD.

Early RGB-D SOD methods mainly utilise the prior knowledge to predict the salient objects by extracting handcrafted features, such as contrast (Cheng et al., 2014), center-surround difference (Zhu and Li, 2018), and boundary background (Feng et al., 2016). However, the handcrafted features are time-consuming and can't represent complex real-world scenarios, which extremely hinder the development of RGB-D SOD. To further boost the performance of RGB-D SOD, convolutional neural network (CNN) is deployed to extract the features from the RGB images and depth maps. Many RGB-D SOD methods (Chen et al., 2020a; Fan et al., 2020a) achieved promising detection results by utilising the feature expression ability of CNN. Existing RGB-D SOD methods usually focus on the fusion and complementarity of the cross-modal features to get better results than adopting a single modality. However, these methods ignore the bi-direction complementarity of RGB and depth features, which can cross-modal information can be used to enhance the features of a single model. However, these methods ignore the bi-direction complementarity of RGB and depth features, which use cross-modal information to enhance the features of a single model.

To deal with the aforementioned problem, we design a joint attention learning network (JALNet), which are is embedded into the JAFM for accomplishing the cross-modal bi-directional complementarity and the joint attention enhance module (JAEM) for capturing the high-level semantic information by the dilated convolution, respectively. The intention of JAFM is to design a bi-directional feature selection and transformation structure to explore effective mutual complementary mechanisms of cross-modal features through the joint cooperation of channel-aware attention and global context-aware attention. The motive of such collaboration is that single-channel attention cannot capture inter-feature dependencies, and a single global context-aware attention fails to effectively represent the associations between channels, so we adopt the combination of the two to remedy the deficiency of a single attention. Furthermore, to enlarge the receptive field, we employ the JAEM to obtain promising features with superior receptive fields by dilated convolution, where the JALNet is employed to filter out the redundant information.

In summary, we have conducted extensive experiments to validate that the proposed JALNet outperforms ten SOTA RGB-D SOD methods over four widely public benchmark datasets. Overall, the main contributions of our work include:

- We propose a JALNet for RGB-D SOD to explore the bi-directional complementarity between RGB and depth modalities. Comprehensive experiments on four public datasets demonstrated that our proposed JALNet outperforms 16 state-of-the-art RGB-D saliency object detection methods.
- We propose a cross-modal joint attention fusion module (JAFM), which consists of channel-aware attention and global context-aware attention, to perfect the cross-modal complementary effect through a joint attention learning mechanism (JALM).
- We design a joint attention enhance fusion module (JAEM) to enlarge the receptive field of the features, which can help the RGB-D SOD models to obtain high-level saliency semantic information for locating the saliency objects and can filter the redundant information by a JALM.

Figure 1 Some visual demonstrations of challenging scenarios, (a) is foreground similar with background (b) is complex background (c) is transparent objects (d) is low-contrast scene (e) is small-object scene (see online version for colours)

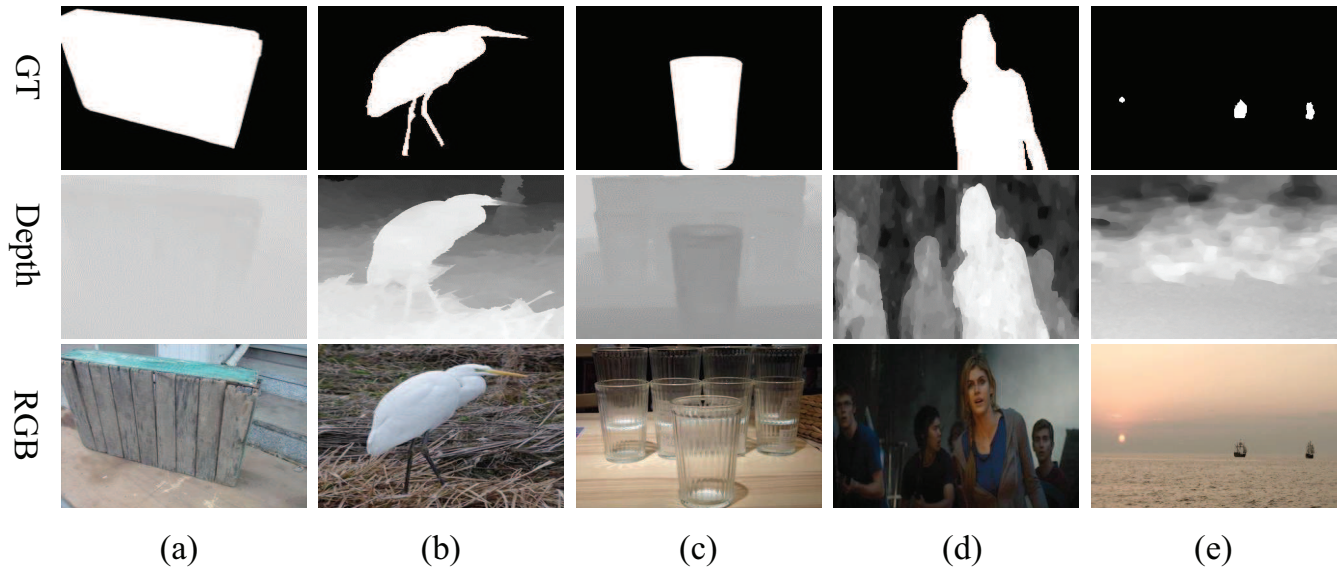
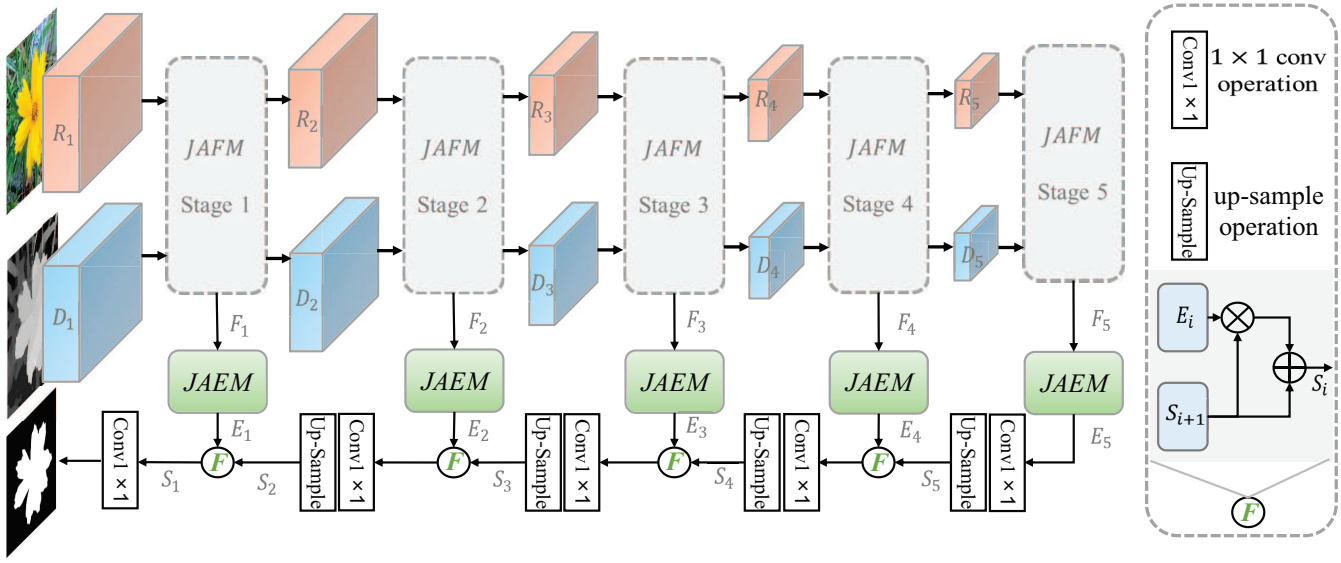


Figure 2 Pipeline of the proposed JALNet. which contains some key stages: cross-modal feature (RGB and depth) encoder, cross-modal JAFM and JAEM (see online version for colours)



Note: The right part shows some annotations of our method.

2 Related work

In this section, we briefly review the current work of RGB saliency object detection and RGB-D saliency object detection, including traditional SOD and deep learning-based SOD.

2.1 RGB SOD

Itti et al. (1998) first proposed the most classical saliency model, which led to a research boom across multiple disciplines, including cognitive psychology, neuroscience, and computer vision. Since then, saliency has received increasing attention from researchers related to computer vision. After Liu et al. (2010) first defined saliency object detection as a binary segmentation problem, SOD has become a computer vision task.

In recent years, with the rapid development of deep learning techniques and their powerful feature extraction capabilities, many researchers have used the CNN to extract features in colour images. Hou et al. (2017) introduced a short-connection to the field of saliency detection, then up-sampled and compressed salient features to output saliency maps by deep supervision. However, deep supervision was time-consuming and prone to overfitting. In Qin et al. (2019a), a residual encoder-decoder structure

is designed to optimise the edges of salient objects. Liu et al. (2021) designed a lightweight SOD model by using a stereo attention network, which consists of a spatial attention and a channel attention. However, the model is too lightweight to have poor generalisation ability. However, they have struggled to deal with challenging scenarios, such as low-contrast environments, similar between foreground and background, and complex backgrounds. To address this issue, researchers introduced depth maps to the SOD due to significant geometric structure embedded into the depth maps.

2.2 RGB-D SOD

Compared with the RGB SOD, the RGB-D SOD has richer information and complementary effects between RGB image and depth map. Similar to RGB SOD methods, RGB-D SOD also can be divided into traditional (no-deep learning) methods and deep learning-based methods. Since RGB-D SOD models need to fuse multi-modal features of RGB and depth images, many saliency detection methods focus on cross-modal fusion schemes. Chen et al. (2019) proposed a cross-scale cross-modal fusion network to integrate RGB features and depth features, and introduced cross-modal interactions into multiple layers. Chen et al. (2020a) designed a gated multi-modality attention (GMA) module to obtain the relevance of global contextual information. However, the gate mechanism will not be able to provide a properly guide when low-quality depth images are encountered. Pang et al. (2020) designed a dynamic dilate pyramid module (DDPM), which fuses multi-modal features using density connection to obtain richer information. However, the cross-modal feature dynamic fusion would loss lose its function when encountering the negative impact of low-quality depth maps, resulting in the lack of robustness for low-quality depth map scenes. Fu et al. (2020) proposed a joint learning and density collaboration RGB-D SOD model (JL-DCF) using a CNN backbone network with shared weights to extract RGB features and depth features, and a density writing fusion strategy to efficiently learn features of different modalities. Most of the above-mentioned guide the cross-modal feature fusion through the interactive approach, these methods cannot exclude the influence of redundant information and negative features in the fusion process during the interactive fusion.

To salving this problem, this paper proposes a JALNet to guide the cross-modal feature fusion process, using global context-aware attention and channel-aware attention mechanisms to form a JALNet.

3 Methodology

3.1 Overall

The overall framework of the proposed JALNet for RGB-D SOD is shown in Figure 2, which consists of three components, namely two-stream encoders, a cross-modal JAFM, and a JAEM for the decoder. To extract the cross-modal features from RGB images and depth maps, we employ the VGG16 backbone (Han et al., 2018) as the decoder, where the final pool layer and fully connected layers are removed. Besides, we design a feature aggregation module (F operation in Figure 2) to integrate cross-modal fused features from different levels.

On the whole, the network follows an encoder-decoder architecture. The RGB features $\{R_i\}_{i=1}^5$ and depth features $\{D_i\}_{i=1}^5$ can be obtained by two-stream encoders, respectively. To integrate $\{R_i\}_{i=1}^5$ and $\{D_i\}_{i=1}^5$, we propose a JAFM to explore the bi-directional complementarity of cross-modal features by the interaction, selection, and fusion strategies. Then, multi-modality fused features $\{F_i\}_{i=1}^5$ can be generated by JAFM. Besides, we introduce a JAEM to improve the global semantic information for the cross-modal fused features by dilated convolution. After, enhanced features $\{E_i\}_{i=1}^5$ can be obtained. Then, we adopt a progressive aggregation manner to integrate multi-scale features $\{S_i\}_{i=1}^5$. Finally, we use sigmoid activate function to inform the predicted saliency map P. Concretely, the encode of saliency can be defined as:

$$E_i = EM(FM(R_i, D_i)), \tag{1}$$

$$\begin{cases} S_5 = UP(Conv_1(E_5)), \\ S_j = UP(Conv_1(Sifmoid(S_{j+1}) \odot E_j \oplus S_{j+1})), \end{cases} \tag{2}$$

where $j \in \{1, 2, 3, 4\}$; *FM* and *EM* mean JAFM and JAEM, respectively; *UP* indicates up-sample operation; $Conv_1$ is 1×1 convolution operation.

3.2 Joint attention fusion module

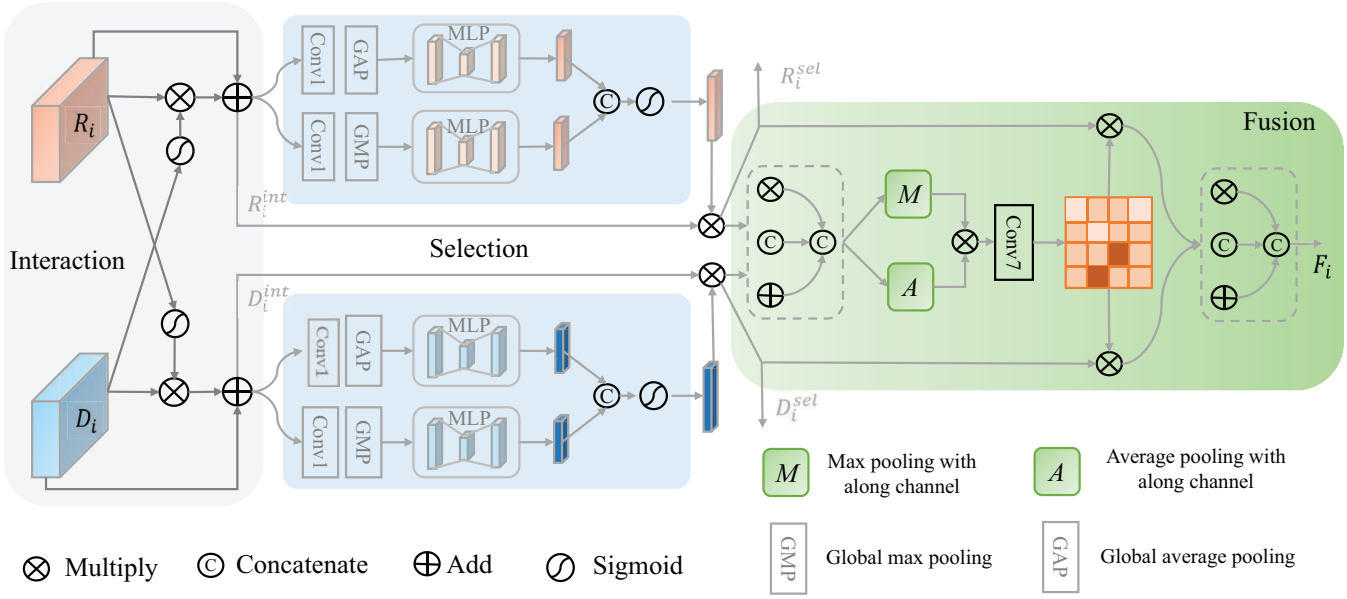
To fuse cross-model features and establish a direct data flow between two-stream encoders and decoders, the JAFM is designed. The specific structure of JAFM is shown Figure 3, which consists of three key stages, namely cross-modal interaction (interaction), channel-based selection (selection), and multi-modality fusion (fusion). The purpose of JAFM is to achieve the perfect cross-modal bi-directional complementary effect, which can adequately take full account of the complementary information of depth features to RGB features and the complementary information of RGB features to depth features. The motivation of bi-directional complementary effect is driven by two situations:

1 depth features can enhance RGB features when dealing with complex scenes

2 RGB features can complement the low-quality depth features.

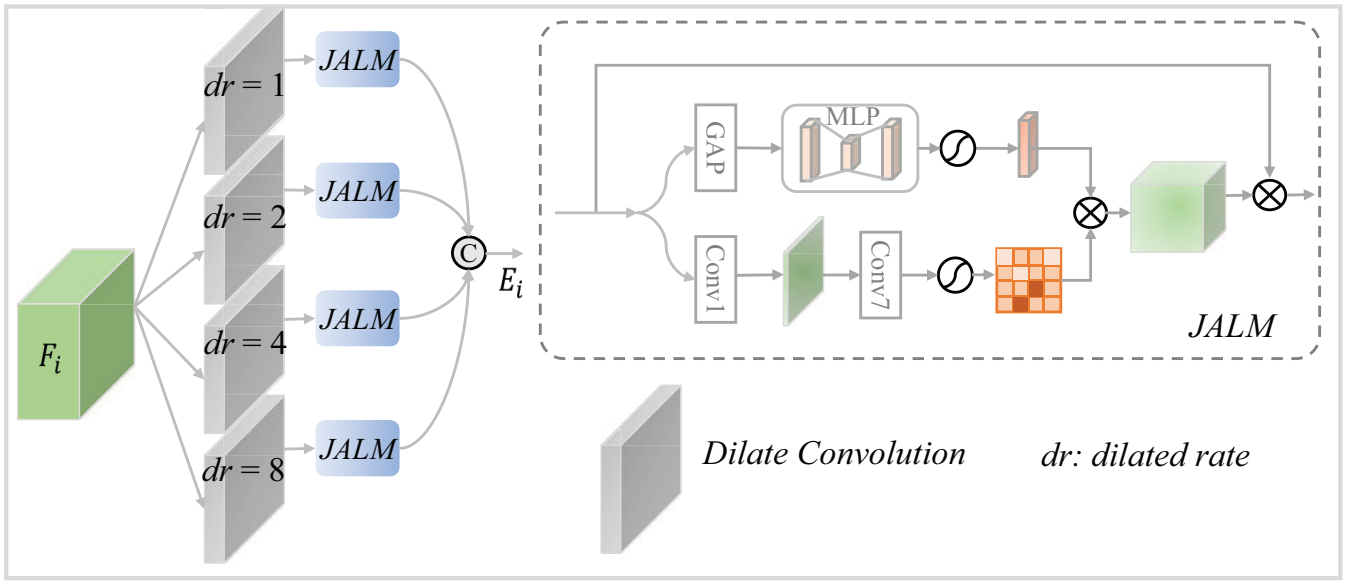
Based on the aforementioned consideration, we design the JAFM to generate the fused features from depth maps and RGB images.

Figure 3 The structure of JAFM, which explores the bi-directional complementary effect by the interaction, selection, and fusion strategies (see online version for colours)



Note: JAFM contains three stages: cross-modal interaction, channel-based selection, and multi-modality fusion.

Figure 4 The structure of JAEM, which consists of JALM and dilated convolution layers (see online version for colours)



Note: The purpose of dilated convolution is to enhance global semantic information of features while the JALM is proposed to improve the discriminative ability of cross-modal features.

3.2.1 *Cross-modal interaction (interaction)*

RGB features and depth features are denoted as $\{R_i\}_{i=1}^{5}$, $\{D_i\}_{i=1}^{5} \in R^{C\times H\times W}$, where i denotes the level of the encoder. C, H, and W indicate the number of channels, the length, and the width of the feature matrix, respectively. Considering the complementarity between the two modalities, we adopt a bi-directional guidance method to achieve cross-modal interaction, where RGB and depth can complete the transmission of information. The process can be described as:

$$R_i^{int} = R_i \odot Sigmoid(D_i) \oplus R_i, \tag{3}$$

$$D_i^{int} = D_i \odot Sigmoid(R_i) \oplus D_i, \tag{4}$$

where R_i and D_i denote the RGB features and the depth features at the i^{th} level in the encoder, and $i \in \{1, 2, 3, 4, 5\}$; $\odot$ and $\oplus$ denote the element-aware multiplication and element-aware addition, respectively; R_i^{int} and D_i^{int} are RGB and depth features after interaction, respectively.

Table 1 Quantitative comparison result on six benchmark datasets, including LFSD, RGBD135, DUT, NLPR, NJU2K, and STERE (see online version for colours)

Dataset	*Metrics*	*DMRA 2019 CVPR*	*CPFP 2019 CVPR*	*TANet 2019 TIP*	*A2dele 2020 CVPR*	*CMWNet 2020 ECCV*	*DANet 2020 ECCV*	*HDFNet 2020 ECCV*	*cmMS 2020 ECCV*	*DFNet 2020 TIP*	*D3Net 2020 TNNLS*	*BiANet 2021 TIP*	*DQSD 2021 TIP*	*DRLF 2021 TIP*	*HAINet 2021 TIP*	*ASIF 2021 T-Cyb*	*ICNet 2021 TIP*	*MoblieSal 2022 TPAMI*	*DCFM 2022 TIP*	*CFID 2022 NCA*	*Ours*
DUT	adpEm ↑	0.930	0.868	–	0.930	0.922	0.929	0.937	0.945	0.853	0.849	0.894	0.889	0.871	0.939	0.883	0.901	0.940	0.951	–	0.958
	adpFm ↑	0.883	0.793	–	0.893	0.865	0.884	0.892	0.906	0.749	0.755	0.810	0.818	0.803	0.906	0.823	0.830	0.912	0.906	–	0.931
	WF ↑	0.857	0.742	–	0.870	0.831	0.846	0.865	0.886	0.698	0.668	0.760	0.775	0.740	0.883	0.779	0.784	0.869	0.889	–	0.914
	MAE ↓	0.048	0.076	–	0.042	0.056	0.047	0.040	0.037	0.104	0.096	0.075	0.072	0.080	0.038	0.072	0.072	0.044	0.035	–	0.028
NLPR	adpEm ↑	0.941	0.924	0.916	0.945	0.940	0.944	0.948	0.947	0.933	0.945	0.939	0.935	0.936	0.952	0.946	0.944	0.953	0.940	0.951	0.957
	adpFm ↑	0.854	0.823	0.795	0.878	0.859	0.865	0.877	0.870	0.838	0.861	0.849	0.842	0.844	0.891	0.871	0.869	0.877	0.854	0.885	0.906
	WF ↑	0.845	0.813	0.779	0.867	0.856	0.849	0.869	0.865	0.827	0.848	0.833	0.843	0.830	0.880	0.856	0.864	0.874	0.856	0.876	0.893
	MAE ↓	0.031	0.036	0.041	0.028	0.029	0.031	0.027	0.027	0.034	0.029	0.032	0.029	0.032	0.025	0.030	0.028	0.025	0.029	0.026	0.022
NJU2K	adpEm ↑	0.920	0.900	0.909	0.916	0.922	0.926	0.933	0.932	0.913	0.915	0.907	0.913	0.903	0.931	0.923	0.912	0.939	0.925	0.929	0.944
	adpFm ↑	0.872	0.837	0.844	0.874	0.880	0.876	0.894	0.886	0.858	0.865	0.848	0.861	0.849	0.896	0.875	0.867	0.894	0.881	0.892	0.909
	WF ↑	0.853	0.828	0.804	0.851	0.856	0.852	0.881	0.867	0.831	0.854	0.811	0.852	0.831	0.879	0.854	0.843	0.874	0.867	0.882	0.894
	MAE ↓	0.051	0.053	0.060	0.051	0.045	0.046	0.037	0.044	0.052	0.046	0.056	0.050	0.055	0.038	0.047	0.052	0.041	0.043	0.038	0.034
STERE	adpEm ↑	0.933	0.907	0.916	0.935	0.930	0.926	0.937	0.937	0.915	0.923	0.925	0.912	0.916	0.937	0.927	0.925	–	0.930	0.933	0.942
	adpFm ↑	0.867	0.830	0.835	0.884	0.869	0.858	0.879	0.879	0.840	0.859	0.869	0.839	0.845	0.890	0.866	0.864	–	0.866	0.879	0.899
	WF ↑	0.850	0.817	0.786	0.867	0.847	0.829	0.863	0.858	0.810	0.837	0.833	0.824	0.821	0.871	0.837	0.843	–	0.849	0.861	0.841
	MAE ↓	0.047	0.051	0.060	0.043	0.043	0.047	0.039	0.043	0.054	0.046	0.050	0.051	0.050	0.038	0.049	0.045	–	0.043	0.043	0.037
RGBD135	adpEm ↑	0.944	0.927	0.919	0.922	0.967	0.960	0.973	–	0.923	0.951	0.925	0.970	0.954	0.967	–	0.959	0.973	0.967	0.943	0.965
	adpFm ↑	0.857	0.829	0.794	0.865	0.900	0.891	0.918	–	0.818	0.870	0.830	0.894	0.868	0.913	–	0.893	0.910	0.896	0.898	0.928
	WF ↑	0.849	0.787	0.738	0.845	0.887	0.848	0.902	–	0.779	0.828	0.774	0.887	0.829	0.897	–	0.867	0.895	0.881	0.875	0.905
	MAE ↓	0.029	0.038	0.046	0.028	0.022	0.028	0.020	–	0.040	0.031	0.038	0.021	0.030	0.019	–	0.027	0.021	0.023	0.023	0.018
LFSD	adpEm ↑	0.899	0.809	0.851	0.880	0.907	0.877	0.882	0.894	0.839	0.863	0.822	0.884	0.872	–	0.861	0.900	0.894	0.905	0.901	0.905
	adpFm ↑	0.849	0.741	0.794	0.835	0.870	0.826	0.830	0.869	0.767	0.804	0.751	0.842	0.821	–	0.827	0.861	0.840	0.861	0.857	0.875
	WF ↑	0.814	0.671	0.071	0.810	0.833	0.789	0.792	0.825	0.070	0.759	0.670	0.796	0.772	–	0.780	0.821	0.800	0.825	0.825	0.841
	MAE ↓	0.075	0.133	0.111	0.073	0.066	0.082	0.085	0.073	0.118	0.095	0.127	0.085	0.089	–	0.090	0.071	0.079	0.068	0.070	0.061

Notes: ↑ and ↓ stand for larger and smaller is better, respectively. The top three results are highlighted in red, violet, and green, respectively.

3.2.2 *Channel-based selection (selection)*

To capture associations between different channel features of R_i^{int} and D_i^{int}, we adopt the global average pooling (GAP), global max pooling (GMP), and multilayer perceptron (MLP) to generate weighted factors representing the channel correlation of R_i^{int} and D_i^{int}. Compared to GAP, GMP can provide unique information about channels that are ignored by GAP. We can compute the enhanced features by GAP and GMP as follows:

$$CA_i^{avg} = MLP(GAP(Conv_1(R_i^{int}))), \tag{5}$$

$$CA_i^{\max} = MLP(GMP(Conv_1(R_i^{int}))), \tag{6}$$

$$R_i^{sel} = R_i^{int} \odot Sigmoid(cat(CA_i^{avg}, CA_i^{\max})), \tag{7}$$

where *GAP* and *GMP* denotes GAP and GAP operation, *MLP* denotes MLP. $Conv_1$ means the 1×1 convolution operation. R_i^{sel} is the RGB feature after channel-based selection. Similar to RGB features, channel-based selective depth feature D_i^{sel} through GAP, MGP, and MLP technologies.

3.2.3 *Multi-modality fusion (fusion)*

The RGB features R_i^{sel} and depth features D_i^{sel} are embedded to multi-modality fusion stage to generate a semantic-aware affine mapping function through spatial attention mechanism, which can capture the degree of variation within features, that is, it can highlight the salient region and suppress the background region. Specially, a cooperation strategy of element-aware addition, element-aware multiplication, and concatenation is adopted to extract significant cues of RGB and depth modalities. Then, a max pooling and an average pooling with along channel be leveraged to capture salient region, which operation can be formulated as:

$$F_i^{raw} = ACM(R_i^{sel}, D_i^{sel}), \tag{8}$$

$$SA_i = Sigmoid(Conv_7(M(F_i^{raw}) \odot A(F_i^{raw}))), \tag{9}$$

where $Conv_7$ denotes the 7×7 convolution operation; M and A are max pooling and an average pooling with along channel; As shown in Figure 3, *ACM* indicates joint use of element-aware addition, element-aware multiplication, and concatenation. SA_i means the attention mask, which can tell which region need attention. Next, we leverage the attention mask SA_i to enhance the RGB feature R_i^{sel} and depth feature D_i^{sel}. Then, using the ACM operation to fuse RGB and depth features, which can be formulated as:

$$F_i = ACM(R_i^{sel} \odot SA_i, D_i^{sel} \odot SA_i). \tag{10}$$

3.3 *Joint attention enhance module*

In order to utilise the cross-modal fused features F_i more rationally and efficiently, we design JAEM with a JALM to enhance the discriminative power and enrich the high-level semantic information of the features. As shown in Figure 4, the JAEM consists of two parts:

1 the dilated convolution layers are employed to enhance the receptive field

2 a JALM is used to improve the discriminative power of cross-modal features.

Firstly, the cross-modal fused features F_i are embedded into the JAEM to generate four features F_i^k with the dilated rate k, where $K \in \{1, 2, 4, 8\}$. The calculation process is shown as follows:

$$F_i^k = DConv_k(F_i),\ i \in \{1, 2, 3, 4, 5\} \tag{11}$$

where $DConv_k$ represents the dilated convolution operation with dilates rates k, $k \in \{1, 2, 4, 8\}$. Then, F_i^k are embedded into the JALM to further enhance the discriminative ability of cross-modal features and eliminate redundant information of cross-modal features, which can be formulated as:

$$CW_i^k = Sigmoid(GAP(MLP(F_i^k))), \tag{12}$$

$$SW_i^k = Sigmoid(Conv_7Conv_1(F_i^k))), \tag{13}$$

where CW_i^k and SW_i^k denote the weight matrices in channel and spatial dimension. Then, CW_i^k and SW_i^k are further combined to generate a joint attention matrix, which is calculated as follows:

$$DF_i^k = F_i^k \odot (CW_i^k \odot SW_i^k), \tag{14}$$

where DF_i^k denotes the refined features with dilated rate k. Finally, we can integrate the enhanced features DF_i^k, where $k \in \{1, 2, 4, 8\}$. The computational procedure is shown as fellow:

$$F_i = cat(DF_i^1, DF_i^2, DF_i^4, DF_i^8).\ i \in \{1, 2, 3, 4, 5\} \tag{15}$$

3.4 Loss function

Similar to works, we adopt the widely used cross-entropy loss (BCE) to supervise our BMCNet. The BCE is computed as:

$$L_{bce} = \frac{1}{H \times W} \sum_{h}^{H} \sum_{w}^{W} [g \log p + (1 - g) \log (1 - p)], \tag{16}$$

where $P = \{p|0 < p < 1\} \in R^{1 \times H \times W}$ and $G = \{g|0, 1\} \in R^{1 \times H \times W}$ represent the predicted value and the corresponding ground truth, respectively. H and W represent the height and width of the input image, respectively. L_{bce} calculates the error between the ground truth G and the predicted P for each position.

4 Experiments

4.1 Datasets and evaluation metrics

To verify the effectiveness of the proposed JALNet, we conduct a comprehensive comparison on five RGB-D benchmark datasets, including LFSD (Li et al., 2014), RGBD135 (Achanta et al., 2009), DUT (Margolin et al., 2014), NLPR (Perazzi et al., 2012), NJU2K (Simonyan and Zisserman, 2014), and STERE (Fan et al., 2020b). LFSD and RGBD135 is a small-scale dataset captured by Kinect camera, which includes 100 and 135 pairs of outdoor and indoor images, respectively. DUT consists of 1200 paired images containing some challenging scenarios, e.g., complex background, transparent object, and low-contrast scenes, which is divided into 800 training samples and 400 testing samples. NLPR contains 1,000 RGB images and corresponding depth maps. Moreover, there are a mass of multi-object scenarios in this dataset. NJU2K contains 1,985 pairs of RGB images and depth maps, where the depth maps are estimated from the stereo images. STERE is fifirst proposed dataset containing 1,000 pairs totally with low-quality depth maps. Follow Hou et al. (2017), we randomly select the 650 samples from NLPR, 1,400 samples from NJU2K, and 800 samples from DUT as the training set, and the remaining samples are classifified as testing set.

To quantitatively analyse the performance of the proposed JALNet, adapt e-measure (adpEm) (Fan et al., 2018), adapt F-measure (adpFm) score (Achanta et al., 2009), weight Fmeasure (WF) score (Margolin et al., 2014), mean absolute error (MAE) (Perazzi et al., 2012), and precision-recall (PR) curve are adopted.

4.2 Implementation details

We use VGG16 (Simonyan and Zisserman, 2014) as the backbone of RGB and Depth encoder. In these two encoders, only the convolutional layer and the final classification layer are retained while the remaining pooling layer and fully connected layer are removed. Our backbone is initialised with ImageNet (Krizhevsky et al., 2012) pre-trained parameter weights. The proposed JALNet is implemented based on PyTorch with a NVIDIA GTX 2080Ti GPU and Adam with momentum optimiser is used to train our model. We set the weight decay 0.1 as batch size as 8, the initial learning rate as $5e^{-5}$, and the momentum as 0.9. We train our model for 80 epochs until convergence.

4.3 Comparison with state-of-the-art methods

To evaluate the performance of the proposed JALNet, we compare our network with other 16 SOTA RGB-D SOD methods, including CPFP (Zhao et al., 2019), HAINet (Li et al., 2021a), ICNet (Li et al., 2021b), DMRA (Piao et al., 2019), TANet (Chen and Li, 2019), A2dele (Piao et al., 2020), CMWNet Li et al. (2020a), DANet (Zhao et al., 2020), HDFNet (Pang et al., 2020), cmMS (Li et al., 2020b), DFNet (Chen et al., 2020b), D3Net (Fan et al., 2020b), BiANet (Zhang et al., 2021), DQSD (Chen et al., 2021), DRLF (Wang et al., 2020), ASIFNet (Li et al., 2021c), MobileSal (Wu et al., 2021), DCFM (Wang et al., 2022), and CFID (Chen et al., 2022). Saliency maps of these methods are generated by the original code under default parameters, or provided by the authors.

Figure 5 Precision-recall curves of our JALNet and other SOD methods on RGBD135, DUT, NLPR, NJU2K, STERE, and LFSD datasets (see online version for colours)

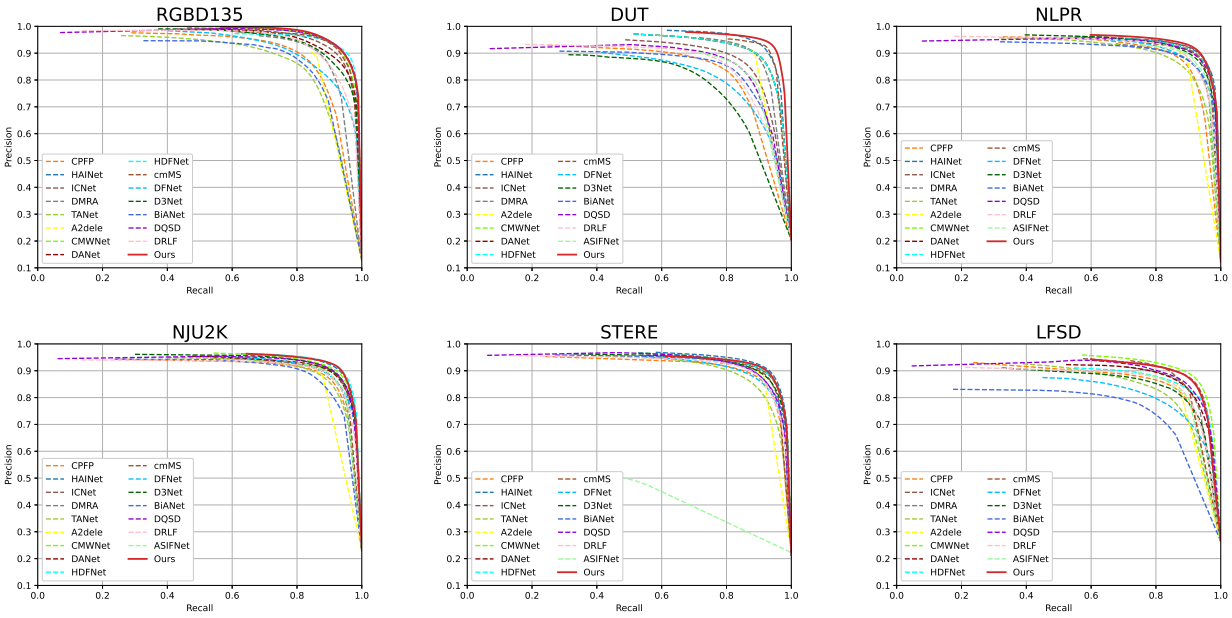


Figure 6 Visual comparison of different RGB-D SOD methods, including, (a) RGB (b) depth (c) GT (d) ours (e) CPFP (f) A2dele (g) CMWN (h) cmMS (i) PGAR (j) D3Net (k) DFNet (l) BiANet (m) ASIF (see online version for colours)

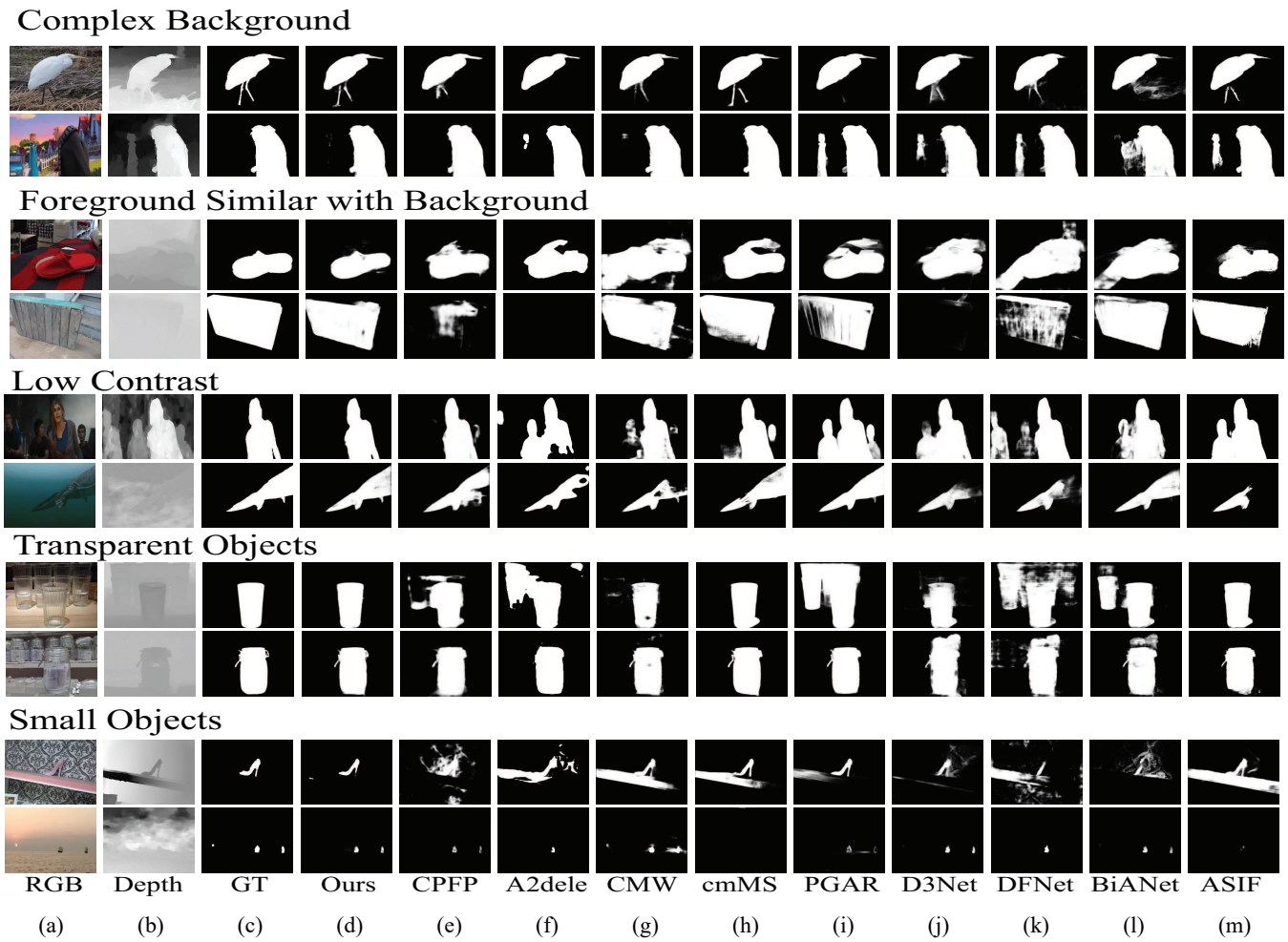


Table 2 Results of ablation study: the contribution of each component of the proposed JALNet (see online version for colours)

Number	Variants	LFSD				NLPR				DUT				STERE			
		adpEm ↑	adpFm ↑	WF ↑	MAE ↓	adpEm ↑	adpFm ↑	WF ↑	MAE ↓	adpEm ↑	adpFm ↑	WF ↑	MAE ↓	adpEm ↑	adpFm ↑	WF ↑	MAE ↓
No. 1	w/o JAFM	0.895	0.861	0.810	0.076	0.950	0.872	0.871	0.026	0.951	0.912	0.890	0.034	0.938	0.886	0.855	0.044
No. 2	w/o JAEM	0.897	0.867	0.829	0.069	0.952	0.892	0.879	0.024	0.954	0.923	0.905	0.031	0.939	0.891	0.866	0.040
No. 3	w/o JAFM + JAEM	0.891	0.857	0.807	0.078	0.945	0.869	0.861	0.028	0.943	0.903	0.879	0.039	0.934	0.880	0.849	0.046
No. 4	Ours	0.905	0.875	0.841	0.062	0.957	0.906	0.893	0.022	0.958	0.931	0.914	0.028	0.942	0.899	0.879	0.037

Notes: The metrics we use are $adpEm$, $adpFm$, WF, and MAE. The best results are highlighted in red.

Table 3 Results of ablation study: the effectiveness of inner strategies in JAFM (see online version for colours)

Number	Variants	LFSD				NLPR				DUT				STERE			
		adpEm ↑	adpFm ↑	WF ↑	MAE ↓	adpEm ↑	adpFm ↑	WF ↑	MAE ↓	adpEm ↑	adpFm ↑	WF ↑	MAE ↓	adpEm ↑	adpFm ↑	WF ↑	MAE ↓
No. 1	JAFM-interaction	0.899	0.852	0.811	0.072	0.942	0.859	0.858	0.028	0.950	0.906	0.887	0.034	0.935	0.879	0.851	0.045
No. 2	JAFM-selection	0.905	0.868	0.832	0.064	0.950	0.873	0.871	0.026	0.951	0.911	0.892	0.034	0.939	0.888	0.871	0.039
No. 3	JAFM-fusion	0.911	0.863	0.822	0.069	0.947	0.868	0.865	0.027	0.954	0.915	0.898	0.032	0.940	0.885	0.859	0.042
No. 4	Ours	0.905	0.875	0.841	0.062	0.957	0.906	0.893	0.022	0.958	0.931	0.914	0.028	0.942	0.899	0.879	0.037

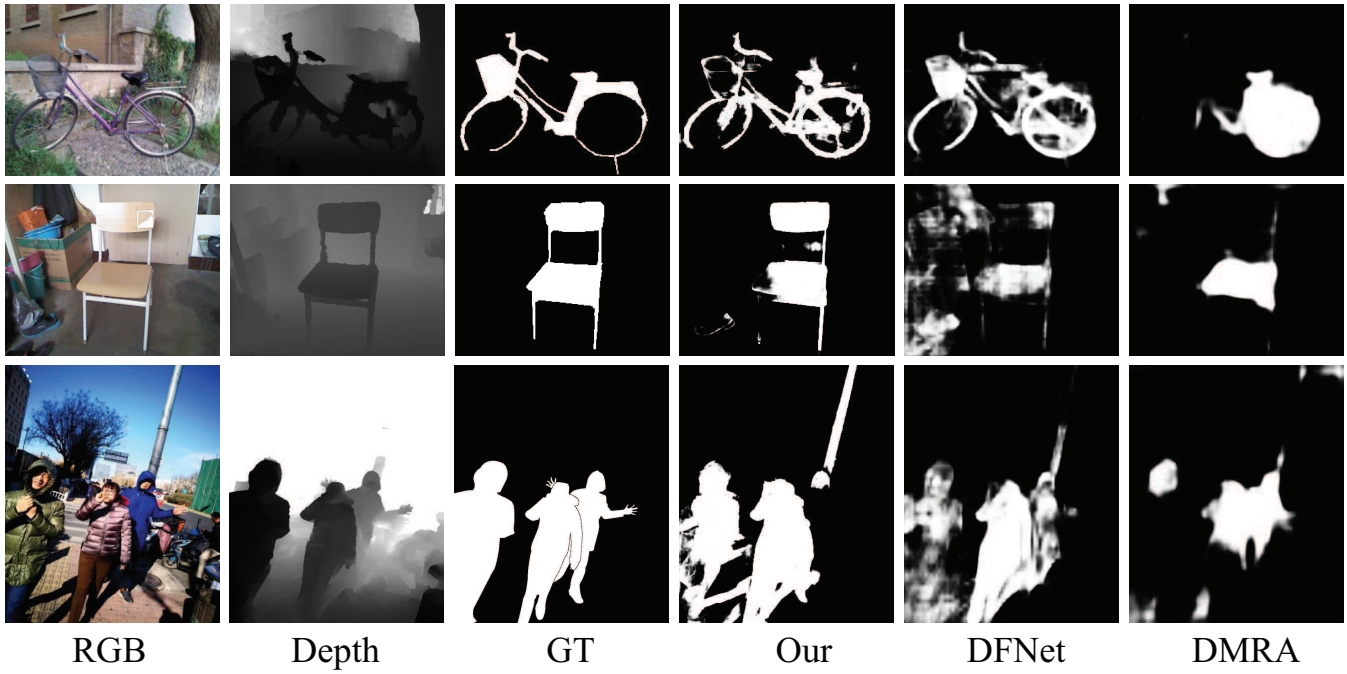
Notes: The metrics we use are $adpEm$, $adpFm$, WF, and MAE. The best results are highlighted in red.

Table 4 Results of ablation study: the effectiveness of inner strategies in JAEM (see online version for colours)

Number	Variants	LFSD				NLPR				DUT				STERE			
		adpEm ↑	adpFm ↑	WF ↑	MAE ↓	adpEm ↑	adpFm ↑	WF ↑	MAE ↓	adpEm ↑	adpFm ↑	WF ↑	MAE ↓	adpEm ↑	adpFm ↑	WF ↑	MAE ↓
No. 1	JAEM-DConv	0.903	0.859	0.826	0.067	0.950	0.872	0.870	0.027	0.953	0.913	0.893	0.034	0.935	0.881	0.853	0.044
No. 2	JAEM-JALM	0.900	0.854	0.807	0.073	0.945	0.858	0.855	0.029	0.942	0.900	0.879	0.037	0.928	0.869	0.838	0.049
No. 3	Ours	0.905	0.875	0.841	0.062	0.957	0.906	0.893	0.022	0.958	0.931	0.914	0.028	0.942	0.899	0.879	0.037

Notes: The metrics we use are $adpEm$, $adpFm$, WF, and MAE. The best results are highlighted in red.

Figure 7 Some failure cases of our JALNet (see online version for colours)



4.3.1 *Quantitative evaluation*

In Table 1, we list the quantitative comparison results of our method and ten RGB-D SOD methods on four public RGB-D SOD datasets including NLPR, RGBD35, SIP and STERE in terms of adpEm, adpFm, WF and MAE metrics. As can see from Table 1, the proposed JALNet performs the best on NLPR, SIP and DES and SSD datasets. On the large-scale SIP dataset, our method can get percentage gain of 2.8%, 2.9%, 4% and 26% in terms of adpEm, adpFm, WF and MAE compare with the suboptimal method D3Net (Li et al., 2021b), which fully illustrates the superior performance of our proposed model on the SIP dataset. On the NLPR dataset, the proposed JALNet reaches optimality

in adpEm, WF and MAE metrics, and the adpFm metric is only 0.001 away from the optimal model.

To further evaluate the performance, the PR curves on six datasets are also reported in Figure 5. It can be note that the proposed JALNet obtains both higher accuracy and recall scores against other comparison state-of-the-art methods on all datasets in Figure 5. This indicated the effectiveness of our JALNet.

4.3.2 Visual comparison

Figure 6 shows qualitative comparison results of the proposed JALNet and different RGB-D SOD methods on various challenging scenarios, including complex background, foreground similar with background, low contrast, transparent objects, and small objects. We compare JALNet with the following methods: CPFP, A2dele, CMWNet, cmMS, PGAR, D3Net, DFNet, BiANet, and ASIF. As shown in Figure 6, the visual results manifest that the proposed JALNet can handle virous challenging scenarios and produce robust prediction of saliency. Taking foreground similar to background scenes as an example. As shown in 2nd row of Figure 6, this scene usually releases some misleading information, which may make some model (e.g., A2dele and D3Net) cannot find any salient objects. However, our method can effectively deal with the case and predict complete salient objects.

4.4 Ablation studies

In this section, we intend to investigate the contribution of each component of the proposed JALNet. To this end, we conduct several ablation studies on LFSD, NLPR, DUT, and STERE datasets, including the effectiveness of JAFM and the importance of JAEM.

4.4.1 The contribution of each component

In our proposed JALNet, the JAFM plays a very important role, which is designed to fuse RGB and depth features. To verify important of the JAFM, we remove the JAFM from JALNet, denoted as w/o JAFM. Similar to JAFM, we remove the JAEM from our method, denoted as w/o JAEM. Besides, we remove JAFM and FAEM, denoted as w/o JAFM + JAEM.

From Table 2, it can be seen that the performance is improved after adding JAFM, which gets the percentage gain of 22.6%, 15.4%, 17.6% and 15.9% in term of MAE score on LFSD, NLPR, DUT, and STERE, respectively. Similar to w/o JAFM, the performance of our method by introducing the JAEM can get 11.3%, 8.3%, 9.7% and 7.5% in terms of MAE score on LFSD, NLPR, DUT, and STERE, respectively. These results adequately prove the effectiveness of JAFM and JAEM. Further, we verify the availability of cooperation of JAFM and JAEM. Comparing with w/o JAFM + JAEM, the proposed JALNet can get significant improvement.

4.4.2 The effectiveness of the JAFM

We introduce the JAFM into our proposed JALNet. To further explore the working mechanism of our proposed JAFM, we further vary the JAFM into three variants:

a we remove the cross-modal interaction, denoted as JAFM-interaction

b removing the channel-based selection, denoted as JAFM-selection

c removing the multi-modality fusion, denoted as JAFM-fusion.

These three sets of comparison experiments on the NLPR, DES datasets are shown in Table 3.

Table 3 presents the compared results of the three variants. From the results, interaction, selection and fusion stages of JAFM are benefited for the proposed JALNet, which manifest that the inner strategies of JAFM are effective. For example, the fusion stage of JAFM can bring a noteworthy improvement, such as WF: 0.841 → 0.822 on LFSD, 0.893 → 0.865 on NLPR, 0.914 → 0.898 on DUT, 0.879 → 0.859 on STERE.

4.4.3 The effectiveness of the JAEM

To demonstrate the contribution of the JAEM, which consists of dilated convolution and JALM, we remove the dilated convolution, denoted as JAEM-DConv. Similar to dilated convolution, removing JALM from FAEM, denoted as FAEM-JALM.

As shown in Table 4, comparing with JAEM-DConv, it can be seen that the dilated convolution generates the percentage gain of 7.5%, 18.5%, 17.6% and 15.9% in terms of MAE on LFSD, NLPR, DUT, and STERE datasets, respectively, which can verify the effectiveness of dilated convolution. Besides, the improvements between JAEM-JALM and the proposed JALNet can confirm that the used JALM is fit to integrate multi-scale features.

4.5 Limitation and analysis

Our proposed JALNet performs very efficiently in most of the SOD tasks, but it encounters some failure cases. As shown in Figure 7, it shows some representative cases. This is because the quality of depth maps for different scenes varies with the scene and depth acquisition device. When we encounter the SOD task with low quality depth maps (such as the first and third rows in Figure 7), it will have great interference and negative impact on the feature fusion process across modalities, which will affect the final generated saliency map. In addition, JALNet will also produce lower quality saliency maps when the salient objects are highly similar to the background in shape and colour (e.g., second row in Figure 7).

5 Conclusions

In this paper, we have developed a novel JALNet for RGB-D SOD to better achieve the cross-modal mutual complementation between the RGB features and the depth features. To this end, a JAFM is designed to integrate cross-modal features by FAM, which effectively uses the mutual complementation of different modality. Besides, to further enhance the semantic information and expression ability of the features, we designed a JAEM to improve the performance of the proposed JALNet. SOD simulates human visual perception system to find the most attractive object in a given scene, and has been widely used in various computer vision tasks, such as image retrieval, visual tracking, medical image segment, and person re-identification. Besides, SOD is one of the most important parts of vision understanding, and it is very important to help machines perceive and understand the captured images for content analysis, such region-of-interest (ROI) extraction in remote sensing image and primary object detection in video.

Acknowledgements

This work was supported by the university-level general projects of Anhui University of science and technology (xjyb2020-04), the National Science Foundation of China (6210071479), the Anhui Natural Science Foundation (2108085QF258), the Natural Science Research Project of Colleges and Universities in Anhui Province (KJ2020A0299), the university-level key projects of Anhui University of Science and Technology (QN2019102), and China Tobacco Henan Industrial Co., Ltd. Science and Technology Projects (AYBW201901).

References

Achanta, R., Hemami, S., Estrada, F. et al. (2009) 'Frequency-tuned salient region detection', *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1597–1604.

Chen, H. and Li, Y. (2019) 'Three-stream attention-aware network for RGB-D salient object detection', *IEEE Transactions on Image Processing*, Vol. 28, No. 6, pp.2825–2835.

Chen, H., Li, Y. and Su, D. (2019) '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.

Chen, Z., Cong, R., Xu, Q. et al. (2020a) 'DPANet: depth potentiality aware gated attention network for RGB-D salient object detection', *IEEE Transactions on Image Processing*, Vol. 30, pp.7012–7024.

Chen, H., Deng, Y., Li, Y., Hung, T-Y. and Lin, G. (2020b) 'RGB-D salient object detection via disentangled cross-modal fusion', *IEEE Transactions on Image Processing*, Vol. 29, pp.8407–8416.

Chen, C., Wei, J., Peng, C. and Qin, H. (2021) 'Depth-quality-aware salient object detection', *IEEE Transactions on Image Processing*, Vol. 30, pp.2350–2363.

Chen, T., Hu, X., Xiao, J., Zhang, G. and Wang, S. (2022) 'CFIDNet: cascaded feature interaction decoder for RGB-D salient object detection', *Neural Computing and Applications*, Vol. 34, No. 10, pp.7547–7563.

Cheng, M-M., Mitra, N.J., Huang, X. et al. (2014) 'Global contrast based salient region detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 3, pp.569–582.

Fan, D-P., Gong, C. and Cao, Y. (2018) 'Enhanced-alignment measure for binary foreground map evaluation', *International Joint Conference on Artificial Intelligence*, pp.698–704.

Fan, D-P., Zhai, Y., Borji, A. et al. (2020a) 'BBSNet: RGB-D salient object detection with a bifurcated backbone strategy network', *European Conference on Computer Vision*, pp.275–292.

Fan, D-P., Lin, Z., Zhang, Z. et al. (2020b) 'Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, No. 5, pp.2075–2089.

Fan, D-P., Ji, G-P., Zhou, T. et al. (2020c) 'PRANet: parallel reverse attention network for polyp segmentation', *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp.263–273.

Feng, D., Barnes, N., You, S. et al. (2016) 'Local background enclosure for RGB-D salient object detection', *IEEE Conference on Computer Vision and Pattern Recognition*, pp.2343–2350.

Fu, K., Fan, D-P., Ji, G-P. et al. (2020) 'JL-DCF: joint learning and densely-cooperative fusion framework for RGB-D salient object detection', *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3052–3062.

Gao, Y., Wang, M., Tao, D. et al. (2012) '3-D object retrieval and recognition with hypergraph analysis', *IEEE Transactions on Image Processing*, Vol. 21, pp.142–149.

Han, J., Chen, H., Liu, N. et al. (2018) 'CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion', *IEEE Transactions on Cybernetics*, Vol. 48, No. 11, pp.3171–3183.

Hong, S., You, T., Kwak, S. et al. (2015) 'Online tracking by learning discriminative saliency map with convolutional neural network', *International Conference on Machine Learning*, pp.597–606.

Hou, Q., Cheng, M-M., Hu, X. et al. (2017) 'Deeply supervised salient object detection with short connections', *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3203–3212.

Itti, L., Koch, C., Niebur, E. et al. (1998) '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.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'Image classification with deep convolutional neural networks', *Advances in Neural Information Processing Systems*, Vol. 25, pp.1097–1105.

Li, N., Ye, J., Ji, Y., Ling, H. and Yu, J. (2014) 'Saliency detection on light field', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2806–2813.

Li, G., Liu, Z., Ye, L., Wang, Y. and Ling, H. (2020a) 'Cross-modal weighting network for RGB-D salient object detection', in *Proceedings of the European Conference on Computer Vision*, Springer, pp.665–681.

Li, C., Cong, R., Piao, Y., Xu, Q. and Loy, C.C. (2020b) 'RGB-D salient object detection with cross-modality modulation and selection', in *Proceedings of the European Conference on Computer Vision*, Springer, pp.225–241.

Li, G., Liu, Z., Chen, M., Bai, Z., Lin, W. and Ling, H. (2021a) 'Hierarchical alternate interaction network for RGB-D salient object detection', *IEEE Transactions on Image Processing*, Vol. 30, pp.3528–3542.

Li, G., Liu, Z. and Ling, H. (2021b) 'ICNet: information conversion network for RGB-D based salient object detection', *IEEE Transactions on Image Processing*, Vol. 29, pp.4873–4884.

Li, C., Cong, R., Kwong, S., Hou, J., Fu, H., Zhu, G., Zhang, D. and Huang, Q. (2021c) 'ASIF-Net: attention steered interweave fusion network for RGB-D salient object detection', *IEEE Transactions on Cybernetics*, Vol. 51, No. 1, pp.88–100.

Liu, T., Yuan, Z., Sun, J. et al. (2010) 'Learning to detect a salient object', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 2, pp.353–367.

Liu, Y., Zhang, X-Y., Bian, J-W. et al. (2021) 'SAMNet: stereoscopically attentive multi-scale network for lightweight salient object detection', *IEEE Transactions on Image Processing*, Vol. 30, pp.3804–3814.

Margolin, R., Zelnik-Manor, L. and Tal, A. (2014) 'How to evaluate foreground maps', *IEEE Conference on Computer Vision and Pattern Recognition*, pp.248–255.

Martinel, N., Micheloni, C. and Foresti, G.L. (2015) 'Kernelized, saliency-based person re-identification through multiple metric learning', *IEEE Transactions on Image Processing*, Vol. 24, No. 12, pp.5645–5658.

Pang, Y., Zhang, L., Zhao, X. et al. (2020) 'Hierarchical dynamic filtering network for RGB-D salient object detection', *European Conference on Computer Vision*, pp.235–252.

Perazzi, F., Krahenb, P.U. et al. (2012) 'Saliency filters: contrast based filtering for salient region detection', *IEEE Conference on Computer Vision and Pattern Recognition*, pp.733–740.

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

Piao, Y., Rong, Z., Zhang, M., Ren, W. and Lu, H. (2020) 'A2dele: adaptive and attentive depth distiller for efficient RGB-D salient object detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.9060–9069.

Qin, X., Zhang, Z., Huang, C. et al. (2019a) 'BET: boundary-aware salient object detection', *IEEE Conference on Computer Vision and Pattern Recognition*, pp.7479–7489.

Qin, X., Zhang, Z., Huang, C., Gao, C. et al. (2019b) 'BASNet: Boundary-aware salient object detection', *IEEE Conference on Computer Vision and Pattern Recognition*, pp.7479–7489.

Simonyan, K. and Zisserman, A. (2014) *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv preprint 434 arXiv:1409.1556.

Wang, F., Pan, J., Xu, S. and Tang, J. (2022) 'Learning discriminative cross-modality features for RGB-D saliency detection', *IEEE Transactions on Image Processing*, Vol. 31, pp.1285–1297.

Wang, X., Li, S., Chen, C., Fang, Y., Hao, A. and Qin, H. (2020) 'Data-level recombination and lightweight fusion scheme for RGB-D salient object detection', *IEEE Transactions on Image Processing*, Vol. 30, pp.458–471.

Wu, Y-H., Liu, Y., Xu, J., Bian, J-W., Gu, Y-C. and Cheng, M-M. (2022) 'MobileSal: extremely efficient RGB-D salient object detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xia, C., Zhang, H., Gao, X. et al. (2020) 'Exploiting background divergence and foreground compactness for salient object detection', *Neurocomputing*, Vol. 383, pp.194–211.

Zhang, Z., Lin, Z., Xu, J., Jin, W-D., Lu, S-P. and Fan, D-P. (2021) 'Bilateral attention network for RGB-D salient object detection', *IEEE Transactions on Image Processing*, Vol. 30, pp.1949–1961.

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

Zhao, X., Zhang, L., Pang, Y., Lu, H. and Zhang, L. (2020) 'A single stream network for robust and real-time RGB-D salient object detection', in *Proceedings of the European Conference on Computer Vision*, Springer, pp.646–662.

Zhu, C. and Li, G. (2018) 'A multilayer backpropagation saliency detection algorithm and its applications', *Multimedia Tools and Applications*, Vol. 77, No. 19, pp.181–197.