Joint Cross-Modal and Unimodal Features for RGB-D Salient Object Detection

Huang, Nianchang; Liu, Yi ; Zhang, Qiang; Han, Jungong

Published in:
IEEE Transactions on Multimedia

Publication date:
2020

Citation for published version (APA):
Huang, N., Liu, Y., Zhang, Q., & Han, J. (Accepted/In press). Joint Cross-Modal and Unimodal Features for RGB-D Salient Object Detection. *IEEE Transactions on Multimedia*.

Joint Cross-Modal and Unimodal Features for RGB-D Salient Object Detection

Nianchang Huang, Yi Liu, Qiang Zhang*, Jungong Han*

***Abstract*—RGB-D salient object detection is one of the basic tasks in computer vision. Most existing models focus on investigating efficient ways of fusing the complementary information from RGB and depth images for better saliency detection. However, for many real-life cases, where one of the input images has poor visual quality or contains affluent saliency cues, fusing cross-modal features does not help to improve the detection accuracy, when compared to using unimodal features only. In view of this, a novel RGB-D salient object detection model is proposed by simultaneously exploiting the cross-modal features from the RGB-D images and the unimodal features from the input RGB and depth images for saliency detection. To this end, a Multi-branch Feature Fusion Module is presented to effectively capture the cross-level and cross-modal complementary information between RGB-D images, as well as the cross-level unimodal features from the RGB images and the depth images separately. On top of that, a Feature Selection Module is designed to adaptively select those highly discriminative features for the final saliency prediction from the fused cross-modal features and the unimodal features. Extensive evaluations on four benchmark datasets demonstrate that the proposed model outperforms the state-of-the-art approaches by a large margin.**

***Index Terms*—RGB-D, saliency detection, multi-branch feature fusion and feature selection.**

I. Introduction

SALIENT Object Detection (SOD) is to detect the most attractive region in the scene by imitating human visual mechanism [1]. It has been applied to a variety of computer vision tasks, including object recognition [2], tracking [3] and segmentation [4], [5], *etc.* Until now, tremendous efforts have been made to detect the salient object in a given image [6], [7], [8], [9], [10], [11], [1], [12]. The earlier methods [11], [1], [12] mainly rely on various types of handcrafted features (e.g., color, intensity and texture) for saliency detection. Recently, with the rapid development of Convolutional Neural Networks (CNNs) [13], [14], [15], [16], [17], CNNs based SOD models [6], [7], [8], [9], [10], [18], [19], [20] have attracted more attention and has achieved significant improvements than conventional models [1], [12].

However, most of these SOD models are designed for visible light images of Red, Green and Blue channels (i.e., RGB images). For some challenging scenarios, for example, as shown in the first row of Fig. 1(a) where the salient object and the background share similar appearance, or as shown in the second row of Fig. 1(b) where the background is complex, these RGB-induced models may be powerless. In order to address such issues, researchers started to look into the possibility of using complementary information acquired by two different cameras to enhance image saliency detection. Fusing RGB and depth (RGB-D) images turns out to be one of the most feasible solutions due to the rapid development of depth sensory technologies, such as Microsoft Kinect [21] and Intel Realsense [22]. Different from RGB images that mainly provide spatial appearances of the scene, depth images provide affluent spatial structures and 3D layout information about the scene, which are robust to light and color changing. Benefiting from the complementary information between RGB-D images, more desirable salient object detection results may be obtained. For example, as shown in the first row of Fig. 1(d), by using the depth information, the salient object in the foreground may be easily distinguished from the background although they have similar spatial appearances. As shown in the second row of Fig. 1(d), multiple objects with similar spatial appearances may also be easily distinguished from each other by using the depth information because they have different distances to the imaging sensor.

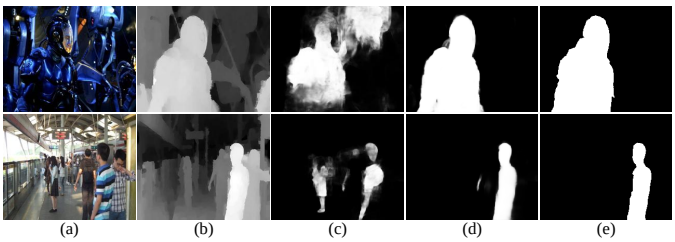


Fig. 1. Illustration of the validity for salient object detection by using RGB-D images. (a) RGB images; (b) Depth images; (c) Saliency maps deduced from RGB images; (d) Saliency maps deduced from RGB-D images; (e) Ground truth. By comparing (c) and (d), it can be easily found that the complementary information between RGB and depth images can boost the saliency detection performance of the traditional RGB-deduced models.

To exploit these complementary information, some CNNs based RGB-D salient object detection models have also been presented in recent years, which can be divided into three categories: pixel-level fusion [23], feature-level fusion [24], [25], [26], [27], [28], [29] and decision-level fusion [30], [31]. In pixel-level fusion, the source RGB-D images are simply considered as four-channel inputs and fed into the networks. In decision-level fusion, two saliency maps are first induced from the input RGB and depth images, respectively and then fused

Nianchang Huang, Yi Liu and Qiang Zhang are with Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an, Shaanxi 710071, China and also with Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China. Email: nchuang@stu.xidian.edu.cn, liuyixd@xidian.edu.cn and qzhang@xidian.edu.cn.

Jungong Han is with Computer Science Department, Aberystwyth University, SY23 3FL, UK. Email: jungonghan77@gmail.com

*Corresponding authors: Qiang Zhang and Jungong Han.

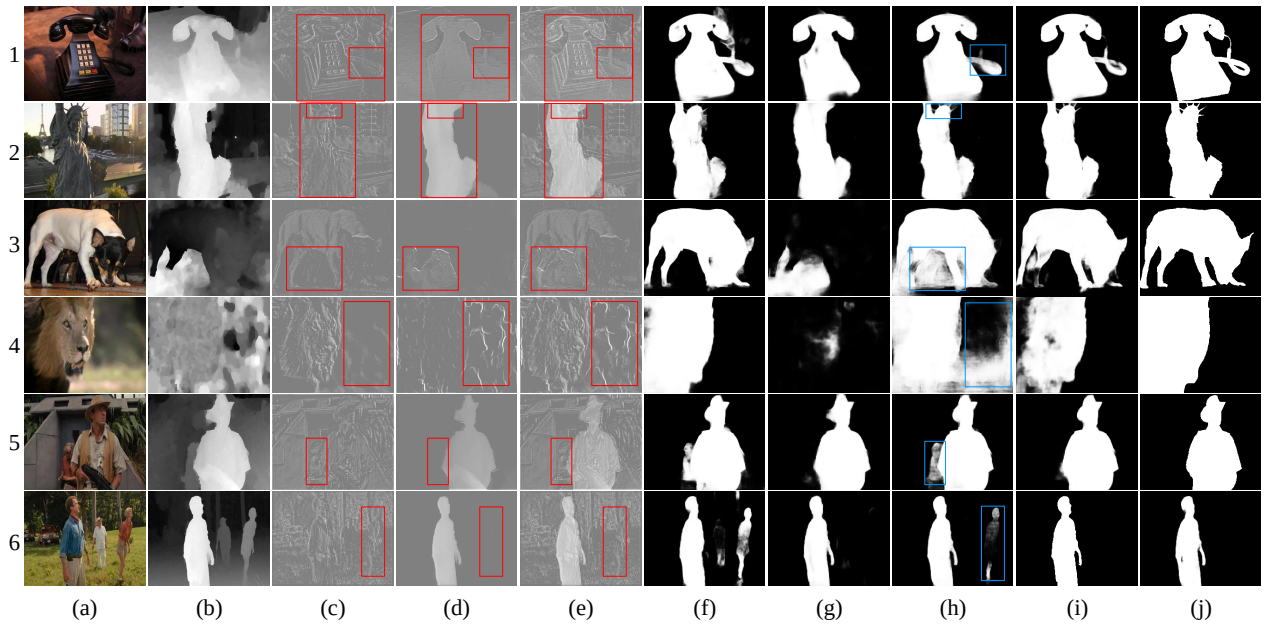


Fig. 2. Illustrations of the discriminative ability of cross-modal features and unimodal features under different cases. (a) RGB images; (b) Depth images; (c) RGB features; (d) Depth features; (e) Cross-modal features; (f) Saliency maps deduced by RGB features; (g) Saliency maps deduced by depth features; (h) Saliency maps deduced by cross-modal features; (i) Saliency maps obtained by our model that uses cross-modal features as well as unimodal features; (j) Ground truth. Both the RGB and depth images in the first and second rows contain many discriminative saliency cues, which may be complementary to each other. Thereby, the corresponding cross-modal features contain the fine details of unimodal RGB features and the depth information of unimodal depth features. However, the depth images in the third and fourth rows are low-quality and the RGB images in the fifth and sixth rows contain multiple objects (e.g., persons) with similar spatial appearances, which introduce some disturbing information to the fused cross-modal features (e.g., the red boxes in (c)-(e)). As a result, some of these objects are mistakenly detected as salient ones by just using the fused cross-modal features (e.g., the blue boxes in (h)).

to obtain the final saliency map. While, in feature-level fusion, a two-stream network is first employed to extract the features from the source RGB and depth images, respectively. Then the extracted features from each unimodal image are fed into a fusion network. The saliency map is finally deduced from these fused features. In general, feature-level fusion can obtain better saliency results than pixel-level fusion and decision-level fusion [25] and thus has attracted more attention in recent years.

Moreover, most RGB-D salient object detection methods based on feature-level fusion usually make use of the complementary cross-modal features from RGB-D images to predict the final saliency maps. As illustrated in the first two rows of Fig. 2, better predictions are generally deduced from the fused cross-modal features than those from unimodal features (e.g., RGB features or depth features). However, it is doubtable that the fused cross-modal features always perform better than those unimodal features, especially when one of the input images has poor visual quality or contains affluent saliency cues. For example, as shown in the red-rectangle regions in the last four rows of Fig. 2, some disturbing features from one of the input images (depth images in the 3rd and 4th rows of Fig. 2, or RGB images in the 5th and 6th rows of Fig. 2) are introduced into the fused features and thus weaken the discriminative ability of the fused features for saliency detection. Accordingly, some background regions are mistakenly determined to be salient ones in the final prediction if using the fused cross-modal features.

Alternatively, better saliency detection results may be obtained if the unimodal (RGB or depth) features and the fused cross-modal features can be simultaneously used during the final saliency prediction. Based on this intuition, we present a novel end-to-end CNN architecture for RGB-D salient object detection in this paper. In the proposed network, a Multi-branch Feature Fusion Module (MFFM) is designed, in which the fused cross-modal features between RGB-D images and the unimodal features from RGB and depth images are simultaneously captured and preserved prior to being fed into the prediction sub-network. On top of that, a Feature Selection Module (FSM) based on the channel-wise attention mechanism is designed to adaptively select those features for the final saliency prediction. As shown in Fig. 2, better saliency detection results can be obtained by jointly using the fused cross-modal RGB-D features as well as the unimodal RGB and depth features, rather than using the fused cross-modal RGB-D features only.

In summary, the main contributions of this work are as follows:

1) An end-to-end CNNs based RGB-D salient object detection network is proposed. As the departure from existing models that only consider the fused cross-modal RGB-D features, our model enables to simultaneously use the fused cross-modal RGB-D features and the unimodal (RGB and depth) features for saliency detection.

2) An MFFM is presented to effectively capture the cross-modal complementary information between RGB-D images, as well as the unimodal features from the RGB images and the depth images. By cascading several MFFMs, the extracted

cross-modal and unimodal features are organized in a coarse-to-fine way and contribute interactively to saliency prediction.

3) Based on the channel-wise attention mechanism, an FSM is designed, where the global and local information are simultaneously employed to adaptively select highly discriminative cross-modal and unimodal features for more accurate salient object detection.

The rest of this paper is organized as follows. In Section II, we briefly introduce previous works related to RGB and RGB-D salient object detection. In Section III, the details of the proposed method are presented. Several experiments are conducted to evaluate the proposed model in Section IV. Finally, Section V concludes this paper.

II. Related Work

A. RGB Salient Object Detection

So far, a number of models have been presented for RGB based salient object detection. Conventional models [11], [1], [12] mainly rely on various types of handcrafted features, such as color, intensity and texture, for saliency detection. Recently, CNNs have been widely used in many computer vision tasks, such as classification [13], [16], [14] and segmentation [32], [33], because of its strong feature learning ability. A lot of CNNs based saliency detection models have also been available [34], [35], [7], [8], [36], [37].

Early CNNs based saliency detection models first employ convolutional layers for feature extraction and then utilize fully connected layers for saliency prediction. For instance, Wang *et al*. [36] first utilized two sub-networks to automatically learn local and global features, given the input images. Then the learned features were fed into multiple fully connected layers for saliency prediction. However, the employed fully connected layers decrease the computational efficiency via dropping the spatial information. Later works address this issue with Fully Convolutional Network (FCN) based saliency detection networks [34], [35], [7], [8], due to the fact that FCN can make dense predictions for pixel-level tasks. Most FCN based salient object detection models exploit multi-level contexts for saliency detection. For example, five levels of features from the VGG-16 network [13] were jointly employed for saliency detection in [35]. In [34], a generic framework aggregating multi-level convolutional features was presented for salient object detection, which simultaneously incorporated coarse semantics and fine details. Furthermore, some works employ multi-scale contextual information to obtain more robust results for the salient objects with different sizes. For example, a multi-scale context-aware feature extraction module was designed in [38], where multiple dilated convolutions were employed to capture multi-scale contextual information for saliency detection.

However, these saliency detection models are merely designed for RGB images. In most cases, these RGB-induced saliency detection models may work well, but they may be powerless for some real-life scenarios, where it is very often that salient objects and backgrounds are similar in appearance or the backgrounds are complex.

B. RGB-D Salient Object Detection

In order to address the above mentioned problems, some works have introduced RGB-D images for saliency detection considering the complementary information within the RGB and depth images. Similar to those RGB-induced salient object detection methods, conventional saliency detection methods for RGB-D images also relied on various types of handcrafted features [39], [40]. For example, a RGB-D based saliency method was presented based on anisotropic center-surround difference in [39]. In [40], based on multi-layer cellular automata, a multi-stage salient object detection framework via minimum barrier distance transform and saliency fusion was proposed for RGB-D images.

In recent years, CNN based RGB-D saliency detection models have become the mainstream [41], [23], [25], [30], [28], [42]. In early CNN based works, the source RGB-D images may be directly considered as the four-channel inputs and fed into a CNN architecture for saliency detection, as in [41], [23]. Lately, various more flexible and complex CNN based RGB-D saliency detection models have been presented to better exploit the cross-modal complementary information. Most of those models employ the fused cross-modal features for saliency detection through involving different multi-modal feature fusion modules [25], [28], [20]. For example, a complementarity-aware fusion module was presented in [25] to effectively exploit the cross-modal complementation as well as the cross-level complementation in the source RGB-D images. In order to better exploit the multi-scale cross-modal features between the source RGB-D images, the depth information was first enhanced by using some contrast priors that had been widely used in the non-deep learning based methods and then was used as an attention map to work with the RGB features for saliency detection via a fluid pyramid integration mechanism in [28].

Meanwhile, other works try to better combine the saliency maps derived from RGB and depth images by generating suitable fusion weights. For instances, in [30], two saliency maps were first generated from the RGB image and the depth image, respectively, by using two independent sub-networks. Then a quality-aware deep neural network was proposed via deep reinforcement learning to model the weights for each source image, by which the two pre-predicted saliency maps were combined to obtain the final saliency map. Similarly, a saliency fusion module was presented to learn a switch map to adaptively fuse the two saliency maps that were pre-deduced from the source RGB and depth images, respectively, via a two-stream CNN in [31].

Recently, in [20], a novel RGB-D Salient Person (SIP) dataset was constructed. Given the SIP dataset and existing six RGB-D datasets, an all-around RGB-D benchmark was presented, in which 31 classical RGB-D salient object detection models were summarized and 17 of them were evaluated. Based on that, a state-of-the-art baseline model, called Deep Depth-Depurator Network (D3Net), was also proposed, which consisted of a depth depurator unit and a feature learning module, performing initial low-quality depth map filtering and cross-modal feature learning, respectively.

In summary, most of these RGB-D saliency detection models mainly focus on how to effectively capture the complementary information within the RGB-D images for saliency prediction. Differently, in this paper, fused cross-modal features and the unimodal features are simultaneously employed for the purpose of performance improvement.

III. Proposed Model

As shown in Fig. 3, the proposed RGB-D salient detection network contains three components: (1) A two-stream subnetwork for unimodal image feature extraction, including one stream for RGB image and the other for depth image; (2) An MFFM for the fusion of cross-modal and cross-level features from the multi-modal RGB-D images as well as the cross-level features from the unimodal RGB and depth images; (3) An FSM based on the attention mechanism to select discriminative features for the saliency prediction. In the following contents, we will discuss these three components in detail, respectively.

A. Two-stream Network for Unimodal RGB and depth Image Feature Extraction

The two-stream unimodal feature extraction network contains two sub-networks with the same structure, which are used to extract the unimodal features from the RGB image and the depth image, respectively. In both sub-networks, the VGG-16 net [13] pre-trained on ImageNet [43] is adopted as the backbone network for fair comparisons with previous works. Other networks, such as Res-Net [16], may also be used. As well, for saliency detection, the last pooling layer and all the full-connected layers are removed from the original VGG-16 for keeping spatial information of input images. For each unimodal RGB or depth image, the modified VGG-16 net provides five levels of features, i.e., Conv 1-2 (containing 64 feature maps of size 224×224, denoted by $\mathbf{F}_i^1$), Conv 2-2 (containing 128 feature maps of size 112×112, denoted by $\mathbf{F}_i^2$), Conv 3-3 (containing 256 feature maps of size 56×56 , denoted by $\mathbf{F}_i^3$), Conv 4-3 (containing 512 feature maps of size 28×28, denoted by $\mathbf{F}_i^4$) and Conv 5-3 (containing 512 feature maps of size 14×14, denoted by $\mathbf{F}_i^5$). Here $i \in \{RGB, depth\}$ denotes the RGB or depth image.

It has been widely proven that multi-scale contextual information is very helpful to salient object detection, since the global context can locate the objects, while the local context can distinguish salient ones from the background [44], [45], [9]. Considering that, an Atrous Spatial Pyramid module with a Residual connection (called as Res_ASPP) is connected to each side-output of the VGG-16 net to capture the multi-scale contextual information of different levels in this paper.

Atrous Spatial Pyramid Pooling (ASPP) was first presented in [46] for semantic segmentation tasks, where four parallel atrous convolutional paths with the same structure but different dilation rates are employed to extract multi-scale contextual information. Recently, it has also been used in some other computer vision tasks, including depth estimation [47] and salient object detection [48]. However, directly adopting ASPP module in our proposed salient object detection model may not work well because of the large dilation rates (e.g., 6/12/18/24) in the original ASPP module [46]. Large dilation rates usually lead to small weights of filters [32]. As the result of that, a reliable contextual relationships among the spatial positions may not be established.

TABLE I
Details of Res_ASPPs for different levels of features.

Level	Dilation rate (r1/r2/r3/r4)	Input channels (N)	Output channels (M)
conv1-2	1/2/3/4	64	64
conv2-2	1/2/3/4	128	128
conv3-3	1/2/3/4	256	192
conv4-3	1/3/5/7	512	256
conv5-3	1/3/5/7	512	384

Therefore, smaller dilation rates are employed in Res_ASPP. Moreover, as shown in Table I, much smaller dilation rates (e.g., 1/2/3/4) are utilized for the shallower levels to capture the local contextual information, while relatively larger dilation rates (e.g., 1/3/5/7) are employed in the Res_ASPPs for the deeper levels to capture the global contextual information. This is mainly due to the fact that shallower levels of features generally contain more spatial details, while deeper levels of features contain more semantics information. Finally, in addition to the four atours convolutional paths, a short connection path with a single regular convolutional layer is added in Res_ASPP as a residual mapping to accelerate the training process [16].

Fig. 4 illustrates the architecture of Res_ASPP. Mathematically, given the m-th level of extracted features $\mathbf{F}_i^m$ from VGG-16 net, the outputs $\widetilde{\mathbf{F}}_i^m$ from the Res_ASPP module are computed by

$$\begin{aligned}\widetilde{\mathbf{F}}_i^m =& \delta(\mathrm{Cat}(\mathrm{AConv}(\mathbf{F}_i^m, \theta_{i,1}^m), \mathrm{AConv}(\mathbf{F}_i^m, \theta_{i,2}^m), \mathrm{AConv}(\\ &\mathbf{F}_i^m, \theta_{i,3}^m), \mathrm{AConv}(\mathbf{F}_i^m, \theta_{i,4}^m)) + \mathrm{Conv}(\mathbf{F}_i^m, \vartheta_i^m)),\end{aligned} \tag{1}$$

where $\delta(*)$ and $\mathrm{Cat}(*)$ denote the ReLU activation function [49] and the concatenation operation, respectively. $\mathrm{AConv}(*, \theta_{i,l}^m)(l = 1, 2, 3, 4)$ refers to the four atrous convolutional layers with the same kernel size of 3×3 but different dilation rates and their corresponding network parameters $\theta_{i,l}^m$. $\mathrm{Conv}(*, \vartheta_i^m)$ denotes a regular convolutional layer with kernel size of 1×1 and its network parameters ϑ_i^m. As discussed above, in addition to the original features $\mathbf{F}_i^m$ from each level of the VGG-16 net, their multi-scale contextual information can also be captured by Res_ASPP. This will greatly benefit the final saliency inference and will be verified in the experimental part.

B. Multi-branch Feature Fusion Module

Given the unimodal features extracted from RGB and depth images, most existing RGB-D saliency detection methods pay more attention to how to fuse these unimodal features [25], [28], [20]. In the subsequent saliency prediction, only the fused cross-modal features are employed and the unimodal features are discarded. This may work well for most cases. However, as discussed in the earlier Section I, in some special

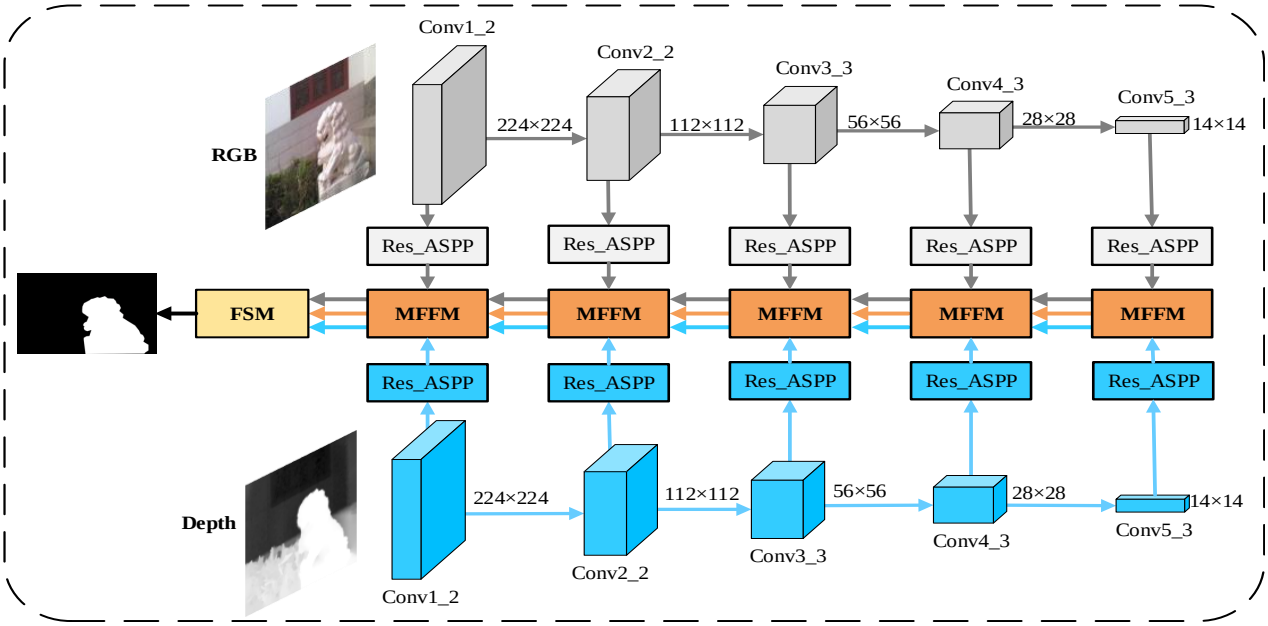


Fig. 3. Diagram of the proposed RGB-D salient object detection network. First, the unimodal RGB and depth features are extracted from the two-stream backbone network. Then, these unimodal features are fed into the Res_ASPP modules to generate more multi-scale unimodal features. Next, these multi-scale unimodal features are fused via the proposed MFFMs to capture cross-modal complementary information between the input RGB-D images. In addition to the fused cross-modal features, the unimodal features from the input RGB and depth images will be also preserved for the saliency prediction via the proposed MFFMs. Finally, those cross-modal features and unimodal features with high discriminability are adaptively selected from the last MFFM for the final saliency prediction by using the proposed FSMs.

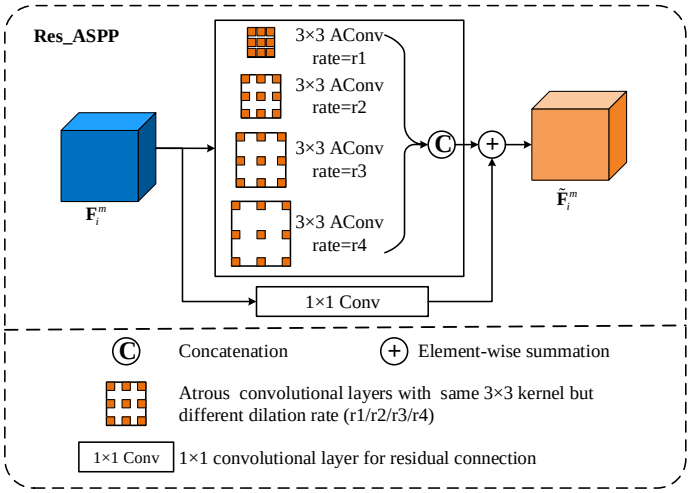


Fig. 4. Architecture of Res_ASPP. In Res_ASPP, four parallel atrous convolutional layers with different dilation rates are first employed to extract multi-scale features. Then, a short connection path is employed as a residual mapping. Finally, the outputs of Res_ASPP are obtained by adding the extracted multi-scale features and the features from the short connection path.

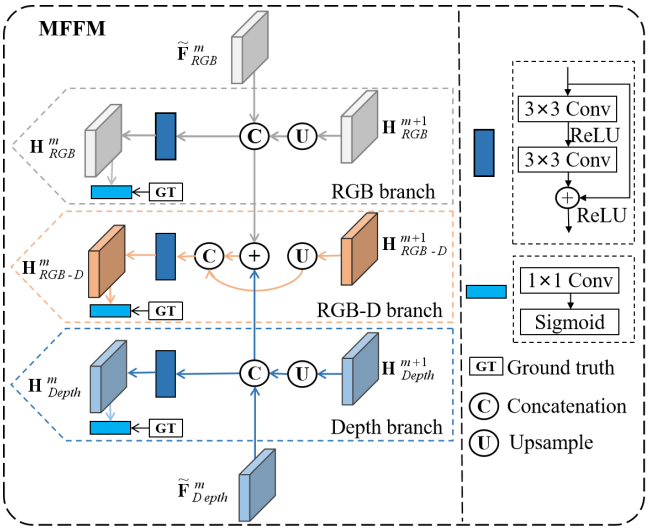


Fig. 5. Diagram of the proposed MFFM. The proposed MFFM contains three branches, one branch for capturing cross-modal complementary information while the other two branches for preserving unimodal (RGB and depth) features to the next stage.

cases, the fused cross-modal features may not always perform better than those unimodal (RGB or depth) features, especially when one of the input images has poor visual quality or contains affluent saliency cues. Only using the fused cross-modal features may not achieve desirable results for these special cases. Alternatively, better saliency detection results may be obtained if the unimodal (RGB and depth) features and the fused cross-modal features are simultaneously used for the final saliency prediction. Considering that, an MFFM is presented to simultaneously preserve the unimodal (RGB and depth) features as well as the fused cross-modal features for the subsequent saliency prediction.

As shown in Fig. 5, the proposed MFFM contains three branches, including one multi-modal branch and two unimodal branches. The multi-modal branch is designed to capture the cross-modal and cross-level complementary features from the multi-modal RGB-D images. The other two unimodal branches are intended to capture the cross-level complementary features

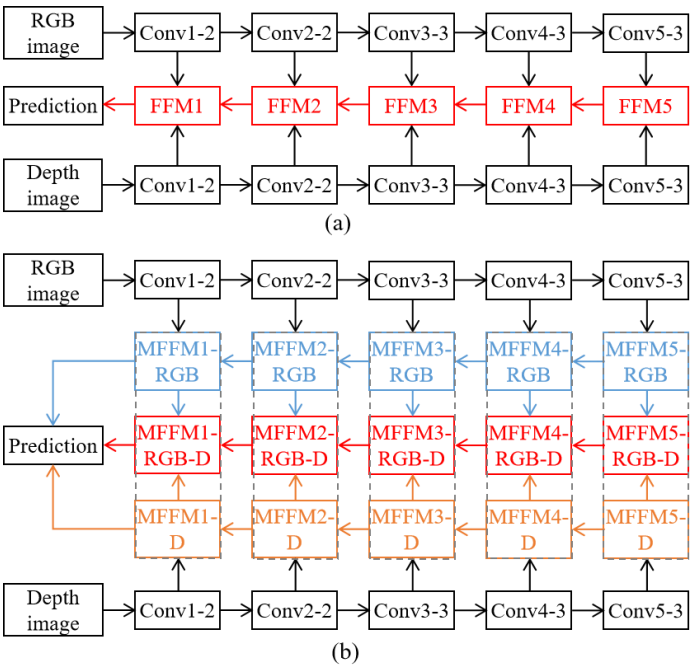


Fig. 6. Difference between the existing FFM and our proposed MFFM. (a) Simplified architecture of some existing RGB-D salient object detection models based on FFM; (b) Simplified architecture of our proposed RGB-D salient object detection model based on MFFM.

from the unimodal RGB or depth image, respectively.

The two unimodal branches (i.e., the RGB branch and the depth branch) in the MFFM share the same structure. Specifically, given the features $\widetilde{\mathbf{F}}_i^m$ ($i \in \{RGB, depth\}$) from Res_ASPP in the m-th level, the outputs $\mathbf{H}_i^m$ from the unimodal RGB or depth branch in MFFM are obtained as follows. The outputs $\mathbf{H}_i^{m+1}$ from the corresponding branch of MFFM in the (m+1)-th level (if exists) are first upsampled by using the bilinear interpolation and then concatenated with the features $\widetilde{\mathbf{F}}_i^m$ from the Res_ASSP module for the current m-th level. After that, the outputs $\mathbf{H}_i^m$ from the corresponding unimodal branch of MFFM in the current m-th level are obtained by performing some convolution and ReLU activation operations on the concatenated features, i.e.,

$$\mathbf{H}_i^m = \begin{cases} \mathrm{RB}(\mathrm{Cat}(\widetilde{\mathbf{F}}_i^m, \mathrm{UP}(\mathbf{H}_i^{m+1})); \omega_i^m), m = 1, 2, 3, 4, \\ \mathrm{RB}(\widetilde{\mathbf{F}}_i^m; \omega_i^m), m = 5, \end{cases} \tag{2}$$

where $\mathrm{RB}(*; \omega_i^m)$ denotes a residual block shown in the right part of Fig. 5, containing two stacked convolutions with a ReLU activation function and a shortcut connection. ω_i^m represents the network parameters for the block $\mathrm{RB}(*; \omega_i^m)$. $\mathrm{UP}(*)$ denotes the upsampling operation with a bilinear interpolation. By doing so, the cross-level complementary information among different levels of features from each unimodal RGB or depth image will be well captured and preserved for the final saliency inference in a coarse-to-fine way.

Different from the unimodal branches that just capture the cross-level complementary features from each unimodal image, the multi-modal branch in MFFM is designed to capture the cross-modal as well as the cross-level complementary features within the multi-modal RGB-D images. For that, as shown in Fig. 5, the unimodal features $\widetilde{\mathbf{F}}_i^m$ from the Res_ASPP module and $\mathbf{H}_i^{m+1}$ from the unimodal branch of MFFM in the coarser level are first concatenated. Then the concatenated features from the RGB branch and the concatenated features from the depth branch are temporally fused via a pixel-wise summation operation. After that, the temporally fused features in the current level and those fused features $\mathbf{H}_{RGB-D}^{m+1}$ from the multi-modal branch of MFFM in the coarser level are further concatenated and fed into a residual block to obtain the final fused features $\mathbf{H}_{RGB-D}^m$. Mathematically, the multi-modal branch can be expressed by

$$\mathbf{H}_{RGB-D}^m = \begin{cases} \mathrm{RB}(\mathrm{Cat}(\mathrm{Cat}(\widetilde{\mathbf{F}}_{RGB}^m, \mathrm{UP}(\mathbf{H}_{RGB}^{m+1})) + \mathrm{Cat}(\widetilde{\mathbf{F}}_{Depth}^m, \mathrm{UP}(\\ \quad \mathbf{H}_{Depth}^{m+1})), \mathrm{UP}(\mathbf{H}_{RGB-D}^{m+1})); \omega_{RGB-D}^m), m = 1, 2, 3, 4, \\ \mathrm{RB}(\widetilde{\mathbf{F}}_{RGB}^m + \widetilde{\mathbf{F}}_{Depth}^m; \omega_{RGB-D}^m), m = 5. \end{cases} \tag{3}$$

By using MFFM, the cross-level complementary features from the unimodal RGB and depth images, together with the cross-modal and cross-level complementary features within the multi-modal RGB-D images, are simultaneously extracted. By cascading several MFFMs, these cross-modal and cross-level features are preserved in a coarse-to-fine way. Besides, inspired by [50] and [51], we also add an intermediate supervision at each branch of MFFM to encourage the cross-modal and cross-level feature fusion timely in each level. This will benefit the final saliency inference greatly.

Fig. 6 illustrates the main difference between the proposed MFFM and the Feature Fusion Module (FFM) used in most of existing RGB-D salient detection models. As shown in Fig. 6(a), existing FFMs, such as Complementarity-Aware Fusion (CA-Fuse) module in [25] and Multi-Modal Feature Fusion network (MMFFnet) in [23], are designed to capture the cross-modal and cross-level complementary features between the multi-modal RGB-D images for the final saliency prediction in a coarse-to-fine way, which is denoted by the red path. In addition to the cross-modal and cross-level complementary features within the multi-modal RGB-D images, the cross-level complementary features within each unimodal RGB or depth image are also extracted and preserved for the final saliency prediction via MFFM. As shown in Fig. 6(b), besides the red path that is used to fuse and transfer the multi-modal features, two extra paths are employed to preserve the unimodal features in MFFM. The blue path and the orange path are designed for the RGB features and the depth features, respectively. In this way, more rich features will be extracted and preserved for the subsequent saliency prediction.

C. Feature Selection Module

MFFM is able to capture the cross-modal and cross-level complementary features from the multi-modal RGB-D images, as well as the cross-level features from the unimodal RGB and depth images. Consequently, it is also ineluctable that some of these cross-modal and unimodal features may contain disturbing information, which will lead to a performance degradation or even wrong predictions. Considering that, as

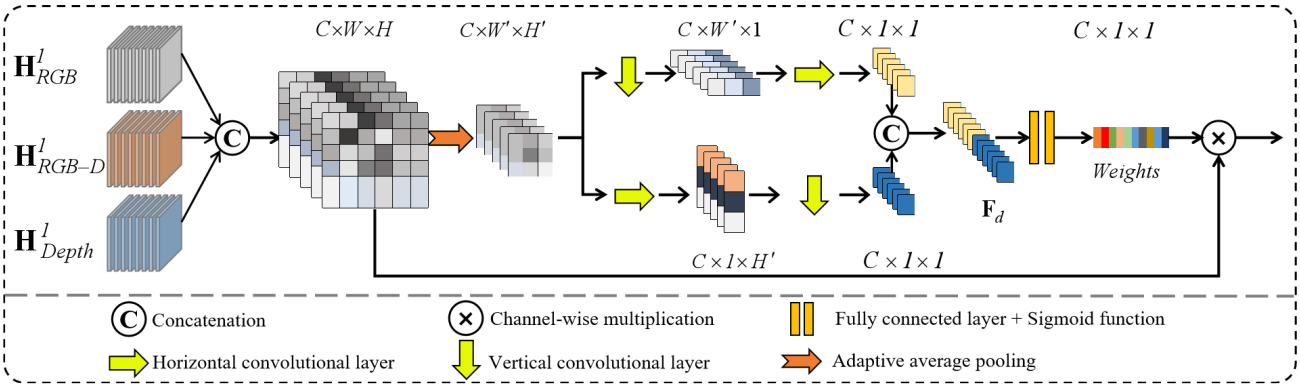


Fig. 7. Architecture of our proposed FSM. First, an adaptive average pooling is employed on concatenated cross-modal and unimodal features to squeeze the spatial information of input features into channel descriptors. Then, these channel descriptors are fed into two parallel branches with mirrored structures for aggregating local information to global information. Finally, weights for different channels of cross-modal and unimodal features are generated by a fully connected layer.

shown in Fig. 7, an FSM based on the channel-wise attention mechanism is presented to further adaptively select highly discriminative cross-modal and unimodal features for the final saliency prediction.

Most existing networks (e.g. Squeeze-and-Excitation Network (SENet) [38] and Convolutional Block Attention Module (CBAM) [52]) carry out the channel-wise attention in two steps. First, a global average pooling is employed to obtain a feature descriptor which contains the global information of each channel. Second, two Fully Connected (FC) layers are employed to fully capture channel-wise dependencies to generate weights for each channel.

However, the global average pooling in those models may capture only global information for each channel of features, but neglects some meaningful local information within each channel of features. For example, as shown in Fig. 8, if only employing global information captured by the global average pooling for selecting highly discriminative features, similar weights may be generated for features in Fig. 8(c) and (d), due to the fact that those features have similar global information, though the features in Fig. 8(c) contain some disturbing local information in the regions marked by red boxes. As a result, inaccurate saliency maps may be deduced (e.g., the local saliency maps in the red boxes of Fig. 8(e)).

Considering that, in addition to global information, local information is also employed to select highly discriminative features in the proposed FSM. Specifically, in the first step of our proposed FSM, an adaptive local average pooling, instead of global average pooling, is employed to squeeze the spatial information of input features into channel descriptors. As shown in Fig. 9, given the input features $\mathbf{X} \in R^{C\times W\times H}$ containing C channels of feature maps of size $W \times H$, each channel of features $x_c \in R^{W\times H}$ in $\mathbf{X}$ are first divided into $W'\times H'$ blocks equally. Then the average pooling is performed on each block and a descriptor $f_c \in R^{W'\times H'}$ for the current channel is obtained.

As shown in Fig. 7, to establish the relations of different channel descriptors with less parameters, the channel descriptors $\mathbf{F} = [f_1, f_2, ..., f_c] \in R^{C\times W'\times H'}$ from the first step of FSM are fed into two parallel branches with mirrored

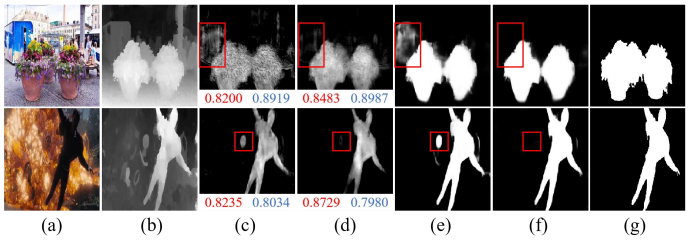


Fig. 8. Illustration of some features selected by jointly using the global and local information (i.e., by employing the proposed FSM) or only by using global information (i.e., by employing the SE block in SENet [38] as an example). (a) RGB images; (b) Depth images; (c) and (d) Some feature maps from MFFM. The red values and blue values are the channel weights generated by FSM and SE block, respectively. (e) Saliency maps deduced by the features from the SE block; (f) Saliency maps deduced by the features from our proposed FSM; (g) Ground truth. SE block tends to generate similar weights for features in (c) and (d) as a result of the following fact. These features in (c) and (d) contain similar global information, although the features in (c) contain some disturbing information in the local regions marked by the red boxes. While, the proposed FSM tends to align higher weights to (c) than those to (d), which owes to the joint local and global information employed by our proposed FSM.

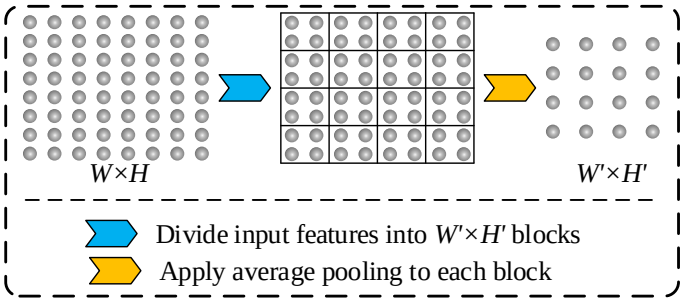


Fig. 9. Illustration of the adaptive average pooling in the proposed FSM.

structures, instead of simply employing a convolutional layer with kernel size of $W' \times H'$, in the second step. In one of the two branches, a horizontal convolutional layer with kernel size of $W' \times 1$ is first employed and then a vertical convolutional layer with kernel size of $1 \times H'$ is employed. Accordingly, in the other branch, a vertical convolutional layer with kernel size of $1 \times H'$ and a horizontal convolutional layer with kernel

size of $W' \times 1$ are sequentially employed. Each convolutional layer in the two branches is followed by the ReLU activation function. By this way, the parameters will be reduced from $C \times W' \times H'$ to $C \times 2 \times (W' + H')$.

Finally, the outputs from the two branches are concatenated and then fed into a FC layer with the Sigmoid function to generate the channel weights $\omega \in R^{C\times 1\times 1}$ for the input features $\mathbf{X}$. After obtaining the channel weights, the final output $\mathbf{Y} \in R^{C\times W\times H}$ of FSM is obtained by multiplying the input features $\mathbf{X}$ with the weights ω in a channel-wise way. In summary, given the features $\{\mathbf{H}^1_{RGB}, \mathbf{H}^1_{Depth}, \mathbf{H}^1_{RGB-D}\}$ from the output of MFFM, FSM can be mathematically expressed by

$$\mathbf{Y} = \mathbf{X} \circ \omega, \tag{4}$$

where $\circ$ denotes the channel-wise multiplication. $\mathbf{X}$ denotes the input features for the proposed FSM, which is constructed by concatenating the features $\{\mathbf{H}^1_{RGB}, \mathbf{H}^1_{Depth}, \mathbf{H}^1_{RGB-D}\}$, i.e.,

$$\mathbf{X} = \mathrm{Cat}(\mathbf{H}^1_{RGB}, \mathbf{H}^1_{Depth}, \mathbf{H}^1_{RGB-D}). \tag{5}$$

The channel-wise weights ω are computed by

$$\begin{aligned}\omega = \sigma(\mathrm{Cat}(\delta(\mathrm{HConv}(\delta(\mathrm{VConv}(\mathrm{AAP}(\mathbf{X}), \gamma_3)), \gamma_2)),\\ \delta(\mathrm{VConv}(\delta(\mathrm{HConv}(\mathrm{AAP}(\mathbf{X}), \gamma_4)), \gamma_5)),), \gamma_1).\end{aligned} \tag{6}$$

Here, $\sigma(*, \gamma_1)$ denotes a FC layer with the Sigmoid function and parameters γ_1. $\mathrm{HConv}(*, \gamma_2)$ and $\mathrm{HConv}(*, \gamma_5)$ are two horizontal convolutional layers with parameters γ_2 and γ_5, respectively. $\mathrm{VConv}(*, \gamma_3)$ and $\mathrm{VConv}(*, \gamma_4)$ are two vertical convolutional layers with parameters γ_3 and γ_4, respectively. $\mathrm{AAP}(*)$ denotes the Adaptive Average Pooling.

By this means, the global information as well as the local information is simultaneously exploited in the proposed FSM to determine whether a feature from MFFM is highly discriminative. Thereby, as shown in Fig. 8, higher weights are assigned to the features in Fig. 8(d) than in Fig. 8(c) by the proposed FSM, due to the fact that the features in Fig. 8(d) contain more accurate saliency cues than those in Fig. 8(c), especially in the local regions marked by red boxes. As a result, the saliency maps deduced by the features from FSM are closer to the ground truth than those deduced by the features from SE blocks.

D. Loss Function

As in [7] and [31], the loss function ζ used to train our network contains two terms, i.e,

$$\zeta = \zeta_{sal} + \zeta_{edge}, \tag{7}$$

where ζ_{sal} denotes the saliency loss to force the predicted saliency map as close to the ground truth as possible. ζ_{edge} denotes the edge-preserving loss to sharp the boundary of the predicted saliency map.

Saliency Loss: As shown in Fig. 5, we use a multi-scale intermediate supervision at each branch of MFFM to encourage the fusion of cross-modal and cross-level features timely in each level. Suppose $\mathbf{S}$ denotes the final saliency map deduced by our proposed method and $\mathbf{S}^m_{RGB}$, $\mathbf{S}^m_{Depth}$ and $\mathbf{S}^m_{RGB-D}$ denote the intermediate saliency maps deduced from the RGB, depth and RGB-D branches of MFFM in the m-th level, respectively. The saliency loss ζ_{sal} is then defined by:

$$\begin{aligned}\zeta_{sal} = \mathrm{L}(\mathbf{S}, \mathbf{Y}) + \sum_m \big(\mathrm{L}(\mathbf{S}^m_{RGB}, \mathbf{Y}^m) + \mathrm{L}(\mathbf{S}^m_{Depth}, \mathbf{Y}^m)\\ + \mathrm{L}(\mathbf{S}^m_{RGB-D}, \mathbf{Y}^m)\big),\end{aligned} \tag{8}$$

where $\mathrm{L}(\mathbf{S}, \mathbf{Y})$ denotes the cross-entropy loss between the saliency map $\mathbf{S}$ and the ground truth $\mathbf{Y}$, i.e.,

$$\mathrm{L}(\mathbf{S}, \mathbf{Y}) = \mathbf{Y}\log(\mathbf{S}) + (1 - \mathbf{Y})\log(1 - \mathbf{S}). \tag{9}$$

Similarly, $\mathrm{L}(\mathbf{S}^m_{RGB}, \mathbf{Y}^m)$, $\mathrm{L}(\mathbf{S}^m_{Depth}, \mathbf{Y}^m)$, $\mathrm{L}(\mathbf{S}^m_{RGB-D}, \mathbf{Y}^m)$ denote the cross-entropy loss between the saliency maps $\mathbf{S}^m_{RGB}$, $\mathbf{S}^m_{Depth}$, $\mathbf{S}^m_{RGB-D}$ and the ground truth $\mathbf{Y}^m$ in m-th level, respectively. $\mathbf{Y}^m$ is sampled from $\mathbf{Y}$ and has the same size as that of $\mathbf{S}^m_{RGB}$, $\mathbf{S}^m_{Depth}$ or $\mathbf{S}^m_{RGB-D}$.

Edge-preserving Loss: To compute the edge-preserving loss ζ_{edge}, two edge maps are first obtained by performing the 'Sobel' operator [1] on the finally predicted saliency map $\mathbf{S}$ and the ground-truth $\mathbf{Y}$, respectively, as suggested in [7] and [31]. Then ζ_{edge} is computed as the sum of the absolute differences between the two edge maps [31], i.e.,

$$\zeta_{edge} = |\mathrm{Sobel}(\mathbf{Y}) - \mathrm{Sobel}(\mathbf{S})|_1, \tag{10}$$

where $\mathrm{Sobel}(\mathbf{Y})$ and $\mathrm{Sobel}(\mathbf{S})$ are the edge maps of $\mathbf{Y}$ and $\mathbf{S}$, respectively, by using 'Sobel' operator. $|\bullet|_1$ denotes the l_1-norm of a matrix and is computed as the sum of the absolute values of all the elements in the matrix.

IV. Experiments

A. Datasets

We conduct several experiments on four publicly available datasets: NJU2000 [53], NLPR [39], STEREO [54] and SIP [20]. NJU2000 [53] contains 2003 stereo RGB-D image pairs with diverse scenarios. NLPR [39] contains 1000 RGB-D image pairs captured by Microsoft Kinect, covering a variety of indoor and outdoor scenes under different illumination conditions. STEREO [54] consists of 797 pairs of binocular RGB-D images. SIP [20] is a newly proposed dataset, which consists of 1000 accurately annotated high-resolution RGB-D image pairs. For a fair comparison, we follow the same data split way as in [6]. Specifically, 1400 samples from NJU2000 and 650 samples from NLPR are randomly selected as the training set. The rest of images are selected as the testing set.

B. Evaluation Metrics

Some standard metrics, including Precision-Recall (PR) curves, F-measure scores, Mean Absolute Error (MAE) and S-measure [55], are employed for performance evaluation. $Precision$ and $Recall$ are computed by comparing the ground

[1] Other edge detection operators (e.g., the traditional image gradient operator, or edge extraction network [51]) may also be employed.

truth and the binarized saliency maps under different thresholds (i.e., from 0 to 255). For MAE, lower values are better and for others, higher values are desirable.

F-measure is a harmonic mean of $Precision$ and $Recall$ and is formulated by:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{11}$$

where, β^2 is set to 0.3, as suggested in [53]. Following [53], max F-measure ($maxF$) and mean F-measure ($meanF$), i.e., max and mean scores of all the F_β values by using different PR pairs, are provided for comparisons.

MAE is computed as the average difference between the predicted saliency map $\mathbf{S}$ and the ground-truth map $\mathbf{Y}$, i.e.,

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\mathbf{S}(x,y) - \mathbf{Y}(x,y)|, \tag{12}$$

where W and H are the width and height of the saliency map (or ground truth), respectively.

S-measure is used to evaluate the structural similarity between the saliency map and the ground truth, which combines a region-aware structural similarity (S_r) and an object-aware structural similarity (S_o), i.e.,

$$\mathrm{S}_{\mathrm{measure}} = \alpha * S_o + (1-\alpha) * S_r, \tag{13}$$

where $\alpha \in [0,1]$ is the balance parameter and is set to 0.5 as default. More details are seen in [55].

C. Implementation Details

We implement our experiments by using the Pytorch [56] toolbox on an NVIDIA 1080Ti GPU. First, the initial parameters of those VGG-16 nets [13] employed for unimodal feature extraction are pre-trained on ImageNet dataset [43]. Other parameters of the proposed model are randomly initialized with Xavier initialization [57]. Then, the entire model is end-to-end trained by employing the Stochastic Gradient Descent (SGD) optimizer with Nesterov momentum [58]. Here, the initial learning rate, weight decay and mini-batch size of the SGD optimizer are set to 0.002, 0.0005 and 4, respectively. Meanwhile, the learning rate will be decayed by a factor of 0.8 in every 30 epochs. During training and testing, all the images are rescaled to the size of 224×224 by employing a bilinear interpolation as in [31].

D. Ablation Experiment and Analysis

In order to verify the validity of these proposed modules in our network, we will perform several ablation experiments on the NJU2000 dataset in this section.

1) Res_ASPP: In order to demonstrate the validity of the proposed Res_ASSP module for the multi-scale contextual feature extraction from the unimodal RGB and depth images, we first remove the Res_ASSP module from our proposed method and obtain another version (w/o Res_ASPP, for short) of our method. After that, two more salient object detection methods (i.e., $w/$ ASPP and $w/$ Res_ASPP, for short) are obtained for comparisons by adding serval ASPP and Res_ASPP modules in w/o Res_ASPP, respectively. In $w/$ ASPP, the dilation rates for all the employed ASPP modules are set to the same values as in [32] (i.e., 6/12/18/24). In $w/$ Res_ASPP (i.e., our proposed method), the dilation rates for the employed Res_ASPP modules are set as in Table I. The quantitative results shown in Table II demonstrate that adding ASPP does not increase and even degrades the saliency detection performance because of the too large dilation rates. In contrast, the salient object detection performance can be greatly improved by using Res_ASPP, which may be due to the smaller dilation rates in Res_ASPP.

TABLE II
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS ON RES_ASPP MODULE.

Methods	$maxF$	$meanF$	$S\text{-}measure$	MAE
w/o Res_ASPP	0.902	0.883	0.895	0.0447
$w/$ ASPP	0.899	0.870	0.897	0.0443
$w/$ Res_ASPP	0.910	0.885	0.903	0.0418

TABLE III
QUANTITATIVE RESULTS BY USING DIFFERENT FUSION MODULES.

Methods	$maxF$	$meanF$	$S\text{-}measure$	MAE
S_RGB	0.868	0.824	0.868	0.0631
S_Depth	0.831	0.802	0.836	0.0808
M_FFM	0.890	0.862	0.892	0.0515
M_MFFM	0.906	0.878	0.900	0.0436

2) MFFM: Several versions of our proposed method (i.e., S_ RGB, S_Depth, M_FFM and M_MFFM, for short, respectively) are provided to test the validity of the proposed MFFM. In S_RGB and S_Depth, only one of the unimodal RGB or depth image is used as the input. In M_FFM and M_MFFM, the multi-modal RGB_D images are used as the input. The only difference between the two methods is that the traditional FFM in Fig. 6(a) is used to fuse the features from the source images in M_FFM, while the proposed MFFM is employed in M_MFFM. As well, for fair comparisons, the proposed FSM is removed from these methods. That is to say, the extracted features from the RGB-D image are directly employed to predict the saliency maps in S_RGB and S_Depth. The features from FFM/MFFM are also directly employed for saliency prediction. Some visual and quantitative results obtained by different methods are shown in Fig. 10 and Table III.

As expected and shown in the first two rows of Fig. 10, both M_FFM and M_MFFM perform better than S_RGB and S_Depth. This demonstrates that the fused cross-modal features from the multi-modal RGB-D images will provide more saliency cues for the saliency detection and thus obtain better saliency detection results than those obtained by the features from one of the unimodal input images in most cases. However, as shown in the last four rows of Fig. 10, when one of the input images has much poor visual quality or contains affluent saliency cues, M_FFM performs worse than S_RGB or S_Depth. This indicates that the fused features from the multi-modal input images may be degraded because

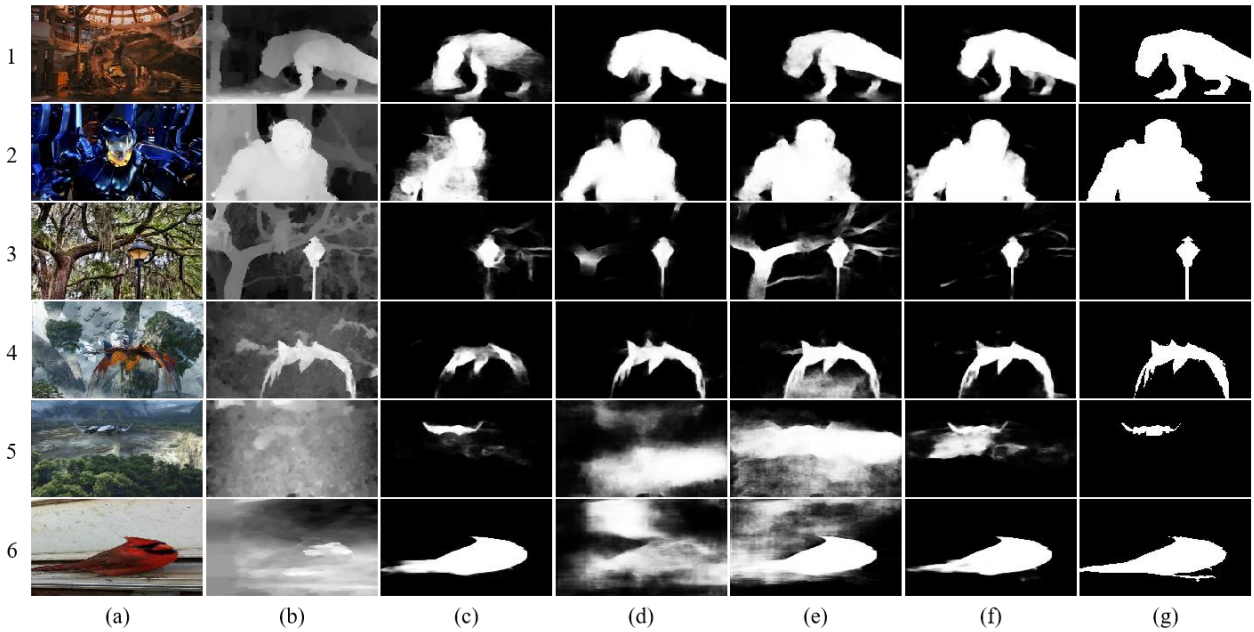


Fig. 10. Illustration of the validity of our proposed MFFM. (a) RGB images; (b) Depth images; (c) Saliency maps obtianed by S_RGB; (d) Saliency maps obtained by S_Depth; (e) Saliency maps obtained by M_FFM; (f) Saliency maps obtained by M_MFFM; (g) Ground truth.

of the features from one of the input images with poor qualities. Accordingly, the saliency detection performance is also degraded by just using the fused features from the multi-modal input images in these special cases. Different from FFM, MFFM preserves the features from each unimodal image as well as those fused features from multi-modal images. As a result, more saliency cues will be provided for the saliency prediction by MFFM. As shown in Fig. 10(f), M_MFFM can still obtain satisfactory results for these special images. The quantitative results of different methods are shown in Table III, which demonstrates the validity of the proposed MFFM for multi-modal image salient object detection again.

3) FSM: As shown in Fig. 11, we illustrate some typical cases to verify the validity of the proposed FSM. In addition, the weights learned by FSM for some channels of features are also provided in Fig. 11 to better demonstrate the validity of the proposed FSM for feature selection.

As shown in the first two rows of Fig. 11, better saliency maps are usually obtained by using the cross-modal features from the RGB-D branch of MFFM in the proposed method than those unimodal features from the RGB or depth branch of MFFM. However, as shown in the last four rows of Fig. 11, the saliency maps deduced by using the features from the RGB-D branch of MFFM are not better and even worse than those deduced by using the features from the RGB or depth branch of MFFM, when one of the input images has much poor visual quality or contains affluent saliency cues.

The weights learned by FSM are also consistent with those visual results. As shown in Fig. 11, the weights assigned to the features from the RGB-D branch are higher than those assigned to the features from the RGB and depth branches for the images in the first two rows. Differently, FSM assigns higher weights to the features from the RGB or depth branch than those from the RGB-D branch for the images in the last four rows. This indicates that FSM may adaptively select those features with higher discriminative ability for the final saliency prediction. As a result, the saliency maps deduced by using those selected features with FSM are very close to the ground truth.

TABLE IV
QUANTITATIVE RESULTS BY USING DIFFERENT ATTENTION MODELS.

Methods	$maxF$	$meanF$	$S\text{-}measure$	MAE
baseline	0.906	0.878	0.900	0.0436
+CBAM	0.901	0.873	0.898	0.0447
+GC block	0.906	0.870	0.900	0.0435
+SE block	0.908	0.878	0.900	0.0433
+FSM	0.910	0.885	0.903	0.0418

As shown in Table IV, we also compare the proposed FSM with some other attention-based modules, including Convolutional Block Attention Module (CBAM) [52], Global Context (GC) block [59] and Squeeze-and-Excitation (SE) block [38]. In Table IV, M_MFFM mentioned above is seen as the baseline method, where FSM is not utilized and the features from MFFM are directly employed to predict the final saliency map. It can be easily found from Table IV that FSM can significantly improve the saliency detection performance of the baseline method. Compared with the other attention modules, FSM may more accurately select those features with highly discriminative ability for the saliency prediction from the outputs of MFFM. This owes to the fact that the local and global information from different channels of features are jointly adopted in FSM to evaluate the importance of each channel.

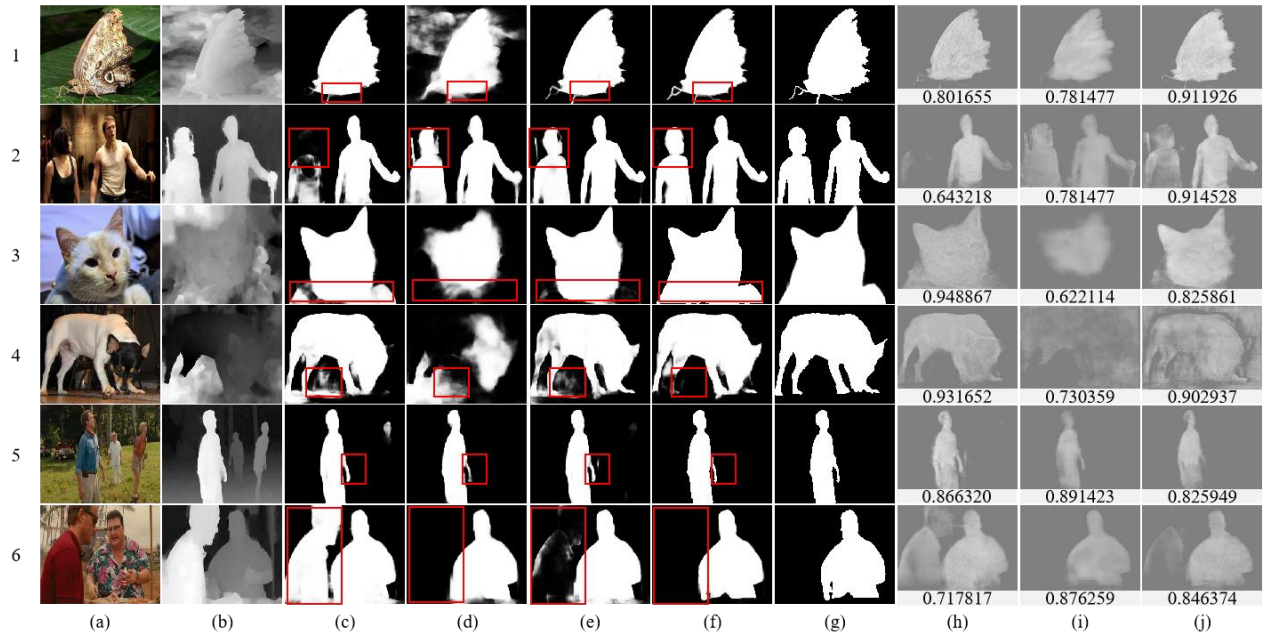


Fig. 11. Illustration of the validity of FSM. (a) RGB images; (b) Depth images; (c) Saliency maps deduced by using the features from the RGB branch of MFFM; (d) Saliency maps deduced by using the features from the depth branch of MFFM; (e) Saliency maps deduced by using the features from the RGB-D branch of MFFM; (f) Saliency maps deduced by using the features from FSM; (g) Ground truth; (h) Features with the highest weights from the RGB branch of MFFM; (i) Features with the highest weights from the depth branch of MFFM; (j) Features with the highest weights from the RGB-D branch of MFFM.

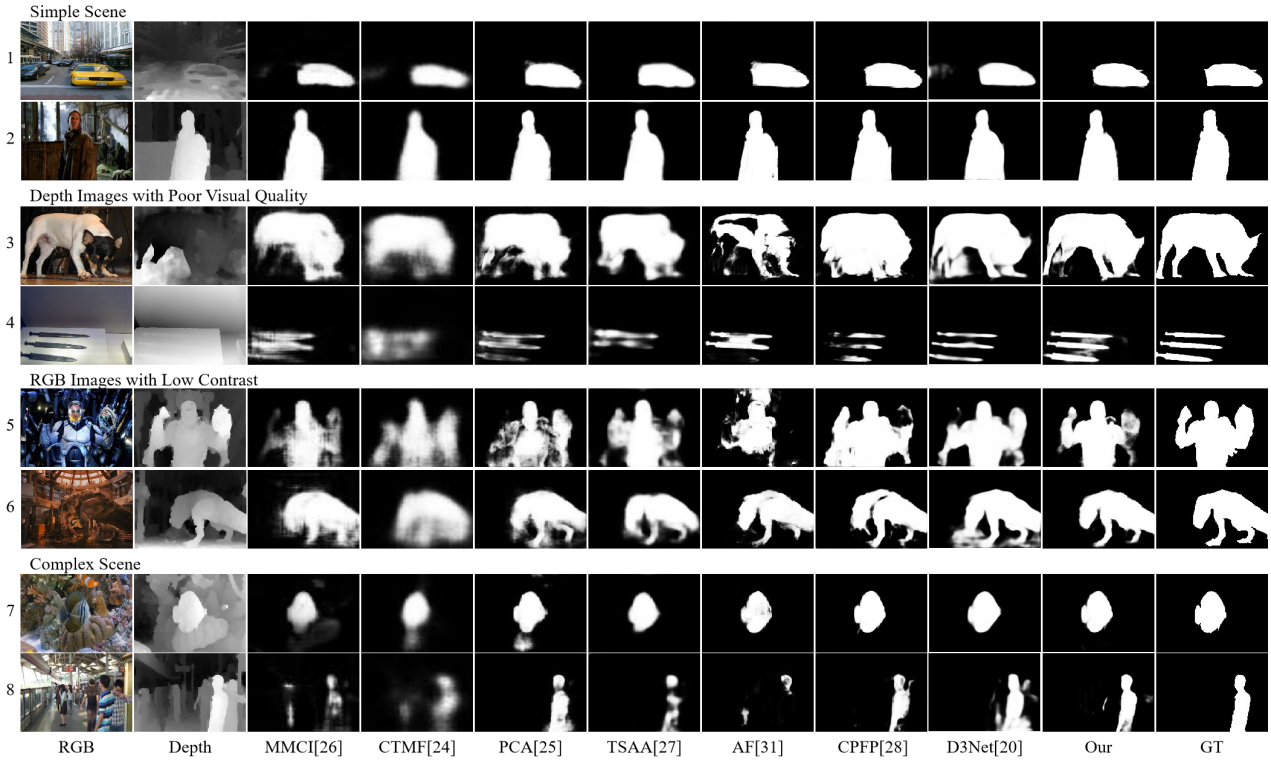
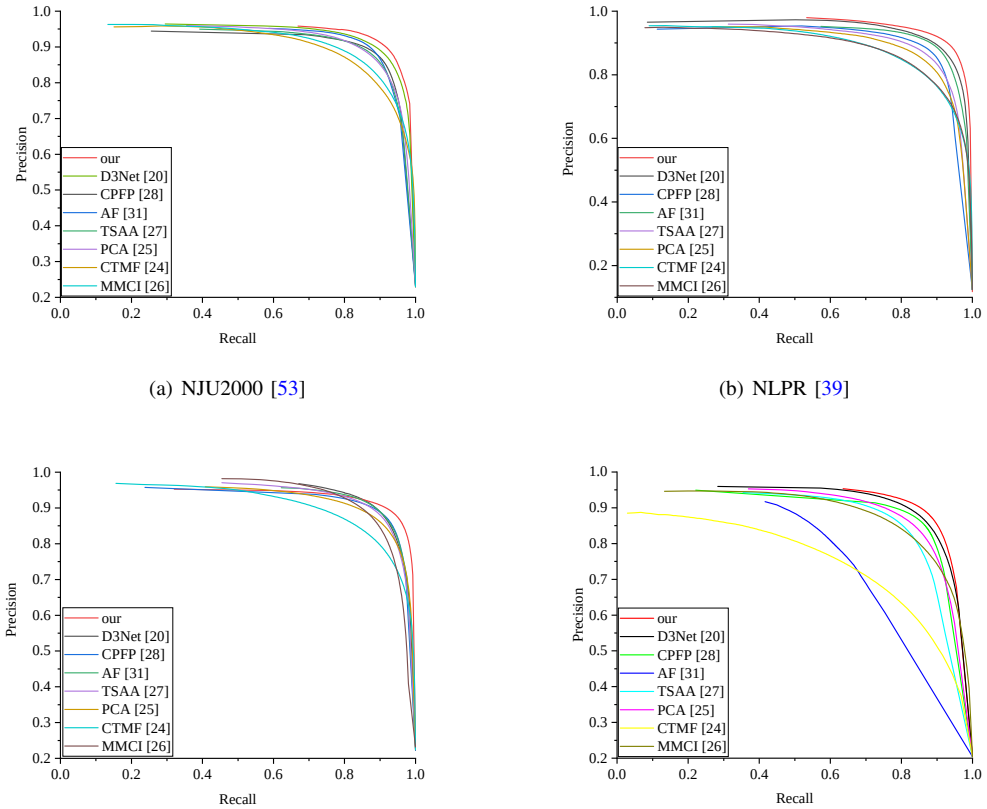


Fig. 12. Some visualization results of different salient object detection methods. As shown in the first two rows, most existing models work well in simple scenes. However, as shown in the rest of rows, most existing works may fail to detect those salient objects under some challenging cases (e.g., depth images with poor visual qualities in the second and third rows, RGB images with low contrast in the fifth and sixth rows or complex scenes in the last two rows), while the proposed model can still obtain good saliency results.

E. Comparison with the State-of-the-Art Models

We compare our model with seven State-Of-The-Art (SOTA) CNNs based RGB-D salient object detection models, including D3Net [20], CPFP [28], AF [31], TSAA [27], PCA [25], CTMF [24] and MMCI [26]. Some visualization results are illustrated in Fig. 12. As shown in the first two rows of



(a) NJU2000 [53] (b) NLPR [39]

(c) STEREO [54] (d) SIP [20]

Fig. 13. PR curves of different methods.

TABLE V

QUANTITATIVE RESULTS BY USING DIFFERENT METHODS. MAX F-MEASURE ($maxF$), MEAN F-MEASURE ($meanF$), S-$measure$ AND MAE ARE EMPLOYED FOR COMPARISONS. FOR $maxF$, $meanF$ AND S-$measure$, HIGHER VALUES ARE DESIRABLE AND FOR MAE, LOWER VALUES ARE DESIRABLE. THE BEST RESULTS ARE SHOWN IN BOLD.

Datasets	Metrics	MMCI [26]	CTMF [24]	PCA [25]	TSAA [27]	AF [31]	CPFP [28]	D3Net [20]	our
NJU2000 [53]	$maxF$	0.867	0.856	0.887	0.887	0.899	0.890	0.903	**0.910**
	$meanF$	0.812	0.784	0.843	0.843	0.869	0.836	0.840	**0.885**
	S-$measure$	0.858	0.849	0.877	0.878	0.880	0.878	0.894	**0.903**
	MAE	0.079	0.085	0.059	0.061	0.053	0.053	0.051	**0.042**
NLPR [39]	$maxF$	0.867	0.856	0.887	0.887	0.899	0.890	0.903	**0.910**
	$meanF$	0.736	0.736	0.801	0.802	0.832	0.830	0.840	**0.870**
	S-$measure$	0.846	0.860	0.873	0.886	0.903	0.888	0.906	**0.920**
	MAE	0.059	0.056	0.044	0.041	0.033	0.036	0.034	**0.027**
STEREO [54]	$maxF$	0.880	0.854	0.884	0.899	0.904	0.897	0.903	**0.910**
	$meanF$	0.857	0.792	0.845	0.878	0.876	0.842	0.844	**0.882**
	S-$measure$	0.869	0.853	0.880	0.886	0.892	0.885	0.895	**0.902**
	MAE	0.074	0.087	0.061	0.055	0.047	0.050	0.053	**0.045**
SIP [20]	$maxF$	0.840	0.719	0.850	0.850	0.771	0.870	0.882	**0.893**
	$meanF$	0.794	0.683	0.825	0.808	0.706	0.818	0.838	**0.854**
	S-$measure$	0.833	0.715	0.834	0.834	0.719	0.850	0.864	**0.873**
	MAE	0.086	0.139	0.075	0.075	0.117	0.064	0.062	**0.056**

Fig. 12, all of the methods mentioned here perform well for those images with simple scenes. However, as shown in the last four rows of Fig. 12, when one of the input images has much poor visual quality or already contains affluent saliency cues, these SOTA methods cannot obtain desirable saliency detection results. For example, some salient objects are not uniformly detected, or parts of the backgrounds are not well suppressed during the saliency detection. As shown in the last

row of Fig. 12, when the backgrounds are much complicated, the salient objects are not even detected by some of these SOTA methods. Differently, our method can still effectively detect the salient objects from these RGB-D images with these challenging scenes. This may be due to the fact that the cross-modal features from the RGB-D images and the unimodal features from the RGB and depth images are simultaneously preserved for saliency prediction in our proposed method. The quantitative results in Table V and Fig. 13 also show that our model significantly outperforms the others in terms of $maxF$, $meanF$, MAE, $S\text{-}measure$ and PR curves.

V. Conclusion

A novel RGB-D salient object detection model has been proposed in this paper, where the cross-level and cross-modal features from the RGB-D image pairs, as well as the cross-level unimodal features from the RGB images and the depth images, are simultaneously captured and preserved during the fusion process by using a proposed MFFM. Furthermore, by virtue of the proposed FSM based on the channel-wise attention mechanism, some channels of features with higher discriminative ability are selectively boosted for the final saliency prediction and some channels of features with less useful information are also suppressed. As a result, when one of the input images has much poor visual quality or contains affluent saliency cues, or when the backgrounds of the scenes are much complex, the proposed method can still effectively detect the salient objects from the scenes. Numerous of experiments have demonstrated the superiorities of the proposed method over the state-of-the-arts.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 61773301 and 61876140 and the China Postdoctoral Support Scheme for Innovative Talents under Grant No. BX20180236.

References

[1] A. Borji, M. Cheng, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational Visual Media*, vol. 5, pp. 117–150, 2014.

[2] Z. Ren, S. Gao, L. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 769–779, 2013.

[3] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *in Proceedings of the International Conference on Machine Learning*, 2015, pp. 597–606.

[4] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5038–5047.

[5] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1742–1756, 2017.

[6] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1734–1746, 2018.

[7] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6593–6601, 2017.

[8] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3127–3135.

[9] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1741–1750.

[10] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2386–2395.

[11] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[12] Y. Liu, J. Han, Q. Zhang, and L. Wang, "Salient object detection via two-stage graphs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1023–1037, 2018.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 630–645.

[15] Q. Xie, O. Remil, Y. Guo, M. Wang, M. Wei, and J. Wang, "Object detection and tracking under occlusion for object-level RGB-D video segmentation," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 580–592, 2017.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

[18] Y. Tang and X. Wu, "Salient object detection using cascaded convolutional neural networks and adversarial learning," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2237–2247, 2019.

[19] K. Fu, Q. Zhao, and I. Y. Gu, "Refinet: A deep segmentation assisted refinement network for salient object detection," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 457–469, 2018.

[20] D. P. Fan, Z. Lin, J. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, and M. Cheng, "Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks," *arXiv preprint arXiv:1907.06781*, 2019.

[21] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, pp. 1318–1334, 2013.

[22] I. Realsense, "Introducing intel realsense lidar camera," Website, 2020, https://www.intelrealsense.com/.

[23] R. Huang, Y. Xing, and Z. Wang, "RGB-D salient object detection by a CNN with multiple layers fusion," *IEEE Signal Processing Letters*, vol. 26, no. 4, pp. 552–556, 2019.

[24] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Transactions on Cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2017.

[25] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

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

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

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

[29] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *Proceedings of the EEE International Conference on Multimedia and Expo*, 2018, pp. 199–204.

[30] X. Wang, T. Sun, R. Yang, C. Li, B. Luo, and J. Tang, "Quality-aware multimodal saliency detection via deep reinforcement learning," *arXiv preprint arXiv:1811.10763*, 2018.

[31] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.

[32] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[33] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 801–818.

[34] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 202–211.

[35] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[36] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[37] Y. Liu, J. Han, Q. Zhang, and C. Shan, "Deep salient object detection with contextual information guidance," *IEEE Transactions on Image Processing*, vol. 29, pp. 360–374, 2019.

[38] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 1, pp. 1–13, 2019.

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

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

[41] R. Shigematsu, D. Feng, S. You, and N. Barnes, "Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2749–2757.

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

[43] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[44] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[45] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3127–3135.

[46] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[47] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[48] A. Kronera, M. Sendena, K. Driessensc, and R. Goebela, "Contextual encoder-decoder network for visual saliency prediction," *arXiv preprint arXiv:1902.06634*, 2019.

[49] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the International Conference on Machine Learning*, 2010, pp. 807–814.

[50] C. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proceedings of the Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[51] S. Xie and Z. Tu, "Holistically-nested edge detection," *International Journal of Computer Vision*, vol. 125, no. 1, pp. 3–12, 2017.

[52] S. Woo, J. Park, J. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[53] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGB-D salient object detection: A benchmark and algorithms," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 92–109.

[54] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 454–461.

[55] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4548–4557.

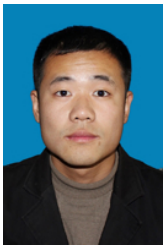
[56] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[57] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[58] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the International Conference on International Conference on Machine Learning*, 2013, pp. 1130–1139.

[59] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," *arXiv preprint arXiv:1904.11492*, 2019.

Nianchang Huang is currently pursuing the Ph.D. degree in School of Mechano-Electronic Engineering, Xidian University, China. His research interests include deep learning and multimodal image processing in computer vision.



Yi Liu received the B. S. degree from Nanjing Institute of Technology, Nanjing, China, in 2012, the M. S. degree from Dalian University, Dalian, China, in 2015, and Ph.D. degree from Xidian University, Xi'an, China, in 2019. He is currently working at Xidian University. He was a visiting student at Lancaster University from September 2018 to September 2019. His current research interests include computer vision and machine learning.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001,2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. He is currently a professor with the Automatic Control Department, Xidian University, China. His current research interests include image processing, pattern recognition.

Jungong Han is currently a Full Professor and Chair in Computer Science at Aberystwyth University, UK. His research interests span the fields of video analysis, computer vision and applied machine learning. He has published over 180 papers, including 40+ IEEE Trans and 40+ A* conference papers.