Pattern Recognition 122 (2022) 108359

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Discriminative unimodal feature selection and fusion for RGB-D salient object detection[☆]

Nianchang Huang[a], Yongjiang Luo[b], Qiang Zhang[a,*], Jungong Han[c]

[a] Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China
[b] The School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China
[c] Computer Science Department, Aberystwyth University, SY23 3FL, UK

ARTICLE INFO

Article history:
Received 23 February 2021
Revised 7 July 2021
Accepted 29 September 2021
Available online 1 October 2021

Keywords:
RGB-D salient object detection
Discriminative unimodal feature selection
Semantic information
Multi-scale cross-modal feature fusion

ABSTRACT

Most existing RGB-D salient object detectors make use of the complementary information of RGB-D images to overcome the challenging scenarios, e.g., low contrast, clutter backgrounds. However, these models generally neglect the fact that one of the input images may be poor in quality. This will adversely affect the discriminative ability of cross-modal features when the two channels are fused directly. To address this issue, a novel end-to-end RGB-D salient object detection model is proposed in this paper. At the core of our model is a Semantic-Guided Modality-Weight Map Generation (SG-MWMG) sub-network, producing modality-weight maps to indicate which regions on both modalities are high-quality regions, given input RGB-D images and the guidance of their semantic information. Based on it, a Bi-directional Multi-scale Cross-modal Feature Fusion (Bi-MCFF) module is presented, where the interactions of the features across different modalities and scales are exploited by using a novel bi-directional structure for better capturing cross-scale and cross-modal complementary information. The experimental results on several benchmark datasets verify the effectiveness and superiority of the proposed method over some state-of-the-art methods.

1. Introduction

Salient Object Detection (SOD) [1] aims to identify the most visually conspicuous objects or regions in a given image. It is an important pre-processing step for a variety of computer vision applications, including image classification [2] and image segmentation [3,4], etc. Despite the profound progress has been achieved by these RGB salient object detection models [5,6], the challenges, such as complex backgrounds and varying illuminations, still hinder their further advances.

Depth images contain affluent spatial structure information and geometrical cues about the scene, which are robust to the light and color changes. It may provide some additional saliency cues to improve the performance of saliency prediction. Meanwhile, the paired RGB and depth (RGB-D) images under the same scene are easily captured by some depth sensors like Microsoft Kinect and Intel RealSense. Therefore, those aforementioned challenges may be overcome by fully utilizing the complementary information in RGB-D images. Considering that, many RGB-D salient object detection models have been proposed and have made significant progress [7].

Recently, many researchers have raised concerns about the quality of the input RGB-D images [8–10]. In theory, better results may be assured by employing the fused cross-modal features as a result of the fact that these fused cross-modal features contain complementary information between the RGB and depth images. In practice, however, it becomes unclear because the qualities of those input RGB/depth images may not be as good as expected, especially for the depth images, due to some factors such as illumination, device capabilities or noise from other devices. Moreover, the discriminative abilities of the unimodal features extracted from the low-quality images are generally trivial, which will degrade the discriminative abilities of subsequently fused cross-modal features, and vice versa. Therefore, in some special cases, the saliency detection results obtained by the fused cross-modal features from the RGB-D images are not necessarily better than those obtained by the unimodal features from one of the input images. For example, as shown in Fig. 1, when one of the input images has poor quality (e.g., the RGB image in the first row and the depth image in the

[☆] This work is supported by the National Natural Science Foundation of China under Grant nos. 61773301. It is also supported by the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University.

[*] Corresponding author.
E-mail addresses: nchuang@stu.xidian.edu.cn (N. Huang), yjluo@mail.xidian.edu.cn (Y. Luo), qzhang@xidian.edu.cn (Q. Zhang).

https://doi.org/10.1016/j.patcog.2021.108359
0031-3203/© 2021 Published by Elsevier Ltd.

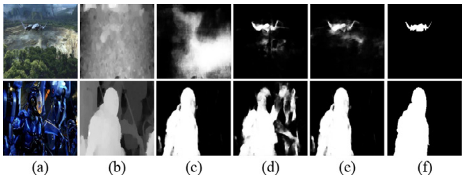

Fig. 1. Illustration of some RGB-D images with low qualities. (a) RGB images; (b) Depth images; (c) Saliency maps deduced from depth images; (d) Saliency maps deduced from RGB images; (e) Saliency maps deduced from RGB-D images; (f) Ground truth. The depth image in the first row and the RGB image in the second row are seen as low-quality ones, considering that the salient objects and backgrounds in the two images have similar depths or share similar appearances.

second row), the saliency maps inferred by the fused cross-modal features are even worse than those inferred from one of the unimodal (RGB/depth) features.

To address this issue, some attention mechanism based unimodal feature selection strategies [8,11,12] are designed to retain the highly discriminative unimodal (RGB/depth) features while discarding the non-discriminative ones. However, these strategies usually pay more attention to the discriminative features that are closely related to salient regions [See Section 4.4.2 for more details]. Differently, we believe that if a local region in the input RGB or depth images provides useful cues to identify salient objects, it is seen as an informative (or high-quality) region for saliency detection, regardless of its saliency. Otherwise, it is seen as a non-informative (or low-quality) one for saliency detection. For example, those regions marked by yellow boxes and those regions marked by red boxes in Fig. 2 may be taken as informative and non-informative ones, respectively. Accordingly, the unimodal features in those informative regions should be retained, while the unimodal features in those non-informative regions should be discarded.

Furthermore, the informative regions generally contain more meaningful semantic information whereas the non-informative regions usually contain more meaningless semantic information. For example, a wing of the airplane can be easily identified in the region marked by the yellow box in the first row of Fig. 2, while nothing can be identified in the region marked by the red box in the third row of Fig. 2. Accordingly, the semantic information extracted from the RGB-D images may also contain some information related to the qualities of the input RGB-D images and thus may be employed to guide the determination of informative regions in the RGB-D images. To implement this intuition, a novel CNNs based RGB-D salient object detection model will be presented in this paper, in which, inspired by attention mechanisms, a Semantic-Guided Modality-Weight Map Generation (SG-MWMG) sub-network is first designed to learn to determine the informative and non-informative (high-quality and low-quality) regions in the input RGB/depth images with the guidance of the semantic information independently extracted from input RGB-D images. By virtue of the proposed SG-MWMG sub-network, the modality-weight maps for the unimodal RGB/depth features are generated at each fusion stage to simultaneously select the discriminative features from those informative regions and suppress the non-discriminative features from those non-informative ones in the subsequent saliency detection.

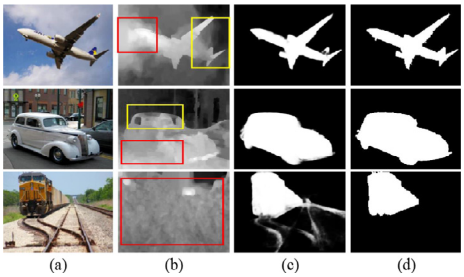

Fig. 2. Visual comparisons of the low-quality regions and the high-quality regions in the RGB-D images. (a) RGB images; (b) Depth images; (c) Saliency maps deduced by the proposed model; (d) Ground truth.

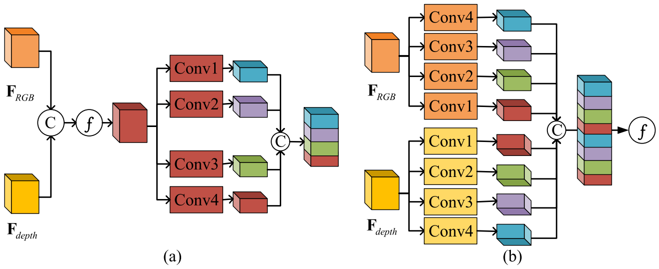

Fig. 3. Architectures of existing multi-scale feature extraction modules. (a) Extracting multi-scale features after multi-modal feature fusion. (b) Extracting multi-scale features before multi-modal feature fusion.

The next step investigates how to extract cross-modal complementary information effectively with the aid of modality-weight maps for better saliency detection. Considering that salient objects have various sizes, shapes and positions, many existing RGB-D SOD models try to extract multi-scale features to handle this issue. However, as shown in Fig. 3, most existing models first extract the features at different scales before or after multi-modal feature fusion and then concatenate the extracted features at different scales for exploiting multi-scale context information. This may give rise to the following two drawbacks. First, the discriminative unimodal features at one scale may be drawn by the features at other scales or other modality. Secondly, they separately extract the features at different scales, which ignores the abundant complementary information across different scales of the features. Considering that, a Bi-directional Multi-scale Cross-modal Feature Fusion (Bi-MCFF) module is then presented in the proposed model to exploit the multi-scale cross-modal complementary information for better saliency detection. By virtue of the proposed Bi-MCFF module, the complementary information across different levels, scales and modalities will be better captured, which will potentially aid in identifying salient objects.

In summary, the main contributions of this paper are as follows.

(1) An RGB-D salient object detection model is presented, where the qualities of input images are considered. This is different from most existing RGB-D salient object detection methods that directly fuse all of the unimodal features without checking the qualities of input images.
(2) An SG-MWMG sub-network is designed to learn to determine the informative and non-informative regions in the input RGB-D images with the guidance of the corresponding semantic information.
(3) An Bi-MCFF module is proposed to capture the cross-scale and cross-modal complementary information from the unimodal features at different scales and levels by using a novel bi-directional structure.

The rest of this paper is organized as follows. We briefly describe some previous works related to the RGB and RGB-D salient object detection in Section 2, followed by the details of our newly proposed method in Section 3. Several experiments are conducted to validate the proposed model in Section 4. Finally, in Section 5, a brief conclusion is made.

2. Related work

2.1. RGB salient object detection models

Earlier RGB salient object detection models mainly rely on various hand-designed features [13–16]. Recently, deep-learning based methods have attracted considerable attention, which have achieved significant improvements compared to those traditional methods, especially since the Fully Convolutional Network (FCN) [17] has been adopted for salient object detection [5,18,19]. Lots of them tried to employ the cross-level contextual information for saliency detection. For example, in Hou et al. [19], an enhanced HED architecture was designed to aggregate the multi-level context information from the deeper layers to the shallower ones by employing multiple short connections. Afterwards, the multi-scale context information was also introduced to help detect the salient objects of different sizes. For example, a Spatial Context-Aware Network (SCA-Net) was designed in Kong et al. [5] to explicitly exploit and assemble global and local contextual information by learning long-path and shortpath dependencies of spatial locations on the feature maps level.

To sum up, although these salient object detection models have achieved profound progress, relying only on RGB images makes them powerless to detect those salient objects under some challenging scenarios, such as for those images with low contrast, complex backgrounds, or for those salient objects sharing similar spatial appearances with the backgrounds.

2.2. RGB-D salient object detection models

A variety of RGB-D salient object detection models have been presented to address those issues mentioned above, by exploiting the complementary information in RGB-D image. A complete survey on RGB-D SOD methods is beyond the scope of this paper and we refer the readers to a recent survey paper [7] for more details. In general, these existing RGB-D salient object detection models can be divided into three categories depending on where they integrate the information: pixel-level fusion, feature-level fusion and decision-level fusion.

In the *pixel-level* fusion based methods [20,21], the RGB and depth images are directly concatenated to form the inputs of four channels for the RGB-D salient object detection models. For example, [22] proposed a conditional probabilistic RGB-D SOD model, which took the concatenated RGB and depth images as inputs and produced several saliency predictions instead of a single saliency map.

For the *feature-level* fusion based RGB-D saliency detection models [8,23–25], the unimodal features of different levels are first extracted from the RGB and depth images, respectively. Then, the extracted unimodal features are fused by employing some specifically designed fusion modules to capture the cross-modal complementary information for better saliency detection. For example, in Hao and Youfu [11], a bottom-up cross-modal distillation stream was designed to explore some new cross-modal representations which might supplement the unimodal features learned from the RGB and depth streams. Then an attention-aware cross-modal cross-level combination (Att-CMCL) block was presented to adaptively select and combine the multi-modal and multi-level features.

With respect to the *decision-level* fusion based methods [26–28], two saliency maps are first independently deduced from the input RGB and depth images, respectively, by using two separate unimodal saliency detection sub-networks. Then the final saliency maps are obtained by fusing the two saliency maps. For example, in Wang and Gong [28], two saliency maps were first predicted from the RGB and depth images by employing two separated sub-networks. Then, those unimodal features from the last layers of those two sub-network were concatenated to generate a weight map by using multiple convolutional layers. The final saliency map can be obtained if fusing the two saliency maps through the generated weight map.

As the departure from most existing RGB-D saliency detection models, the proposed model takes the qualities of input images into account, in which the discriminative unimodal features are selectively fused for saliency detection.

3. Proposed model

As shown in Fig. 4, the proposed RGB-D salient object detection network mainly contains three components: (1) Unimodal feature extraction module to extract the unimodal features from the input RGB-D images; (2) Semantic-Guided Modality-Weight Map Generation (SG-MWMG) sub-network to determine those informative and non-informative regions, given the input RGB/depth images. (3) Bi-directional Multi-scale Cross-modal Feature Fusion (Bi-MCFF) module to extract the cross-modal complementary information at different levels and scales. The three components will be discussed in detail in the following contents, respectively.

3.1. Unimodal feature extraction module

As shown in Fig. 4, the unimodal feature extraction module contains two sub-networks with the same structures. One is used to extract the unimodal features from RGB images and the other is used to extract the unimodal features from depth images. In both of the two sub-networks, the pre-trained VGG-16 net [29] is employed as the backbone network for a fair comparison with previous works. Other networks, such as Res-Net [30], 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 net to maintain more spatial information of the input images. After the input RGB or depth image is fed into the modified VGG-16 net, five levels of features are extracted, i.e., Conv 1–2 (containing 64 feature maps of size 224 × 224, denoted by $\mathbf{F}_m^1$), Conv 2-2 (containing 128 feature maps of size 112 × 112, denoted by $\mathbf{F}_m^2$), Conv 3-3 (containing 256 feature maps of size 56 × 56, denoted by $\mathbf{F}_m^3$), Conv 4-3 (containing 512 feature maps of size 28 × 28, denoted by $\mathbf{F}_m^4$), and Conv 5-3 (containing 512 feature maps of size 14 × 14, denoted by $\mathbf{F}_m^5$). Here $m \in \{RGB, Depth\}$ denotes the RGB or depth image, respectively.

3.2. Semantic-guided modality-weight map generation (SG-MWMG) sub-network

As discussed in Section I, suboptimal saliency detection results may be obtained if the unimodal features are directly fused for saliency prediction without considering the qualities of the input images. Given that, a Semantic-Guided Modality-Weight Map Generation (SG-MWMG) sub-network is presented to determine the qualities of each local region in the input RGB/Depth images by employing the semantic information extracted from the input RGB-D images.

As shown in Fig. 4, the proposed SG-MWMG sub-network adopts an encoder-decoder framework. In the encoder stage, the semantic information is extracted from the RGB-D images by employing an independent sub-network. Specifically, the unimodal features $\mathbf{F}_{RGB}^1$ and $\mathbf{F}_{depth}^1$ are first concatenated as the inputs of the SG-MWMG sub-network. The experimental results show that this will obtain better results than directly taking the concatenated RGB-D images as the inputs. Then, four levels of features (i.e., $\mathbf{H}^i, i = 2, 3, 4, 5$) are thus extracted from the concatenated features by using four stacked convolutional blocks.

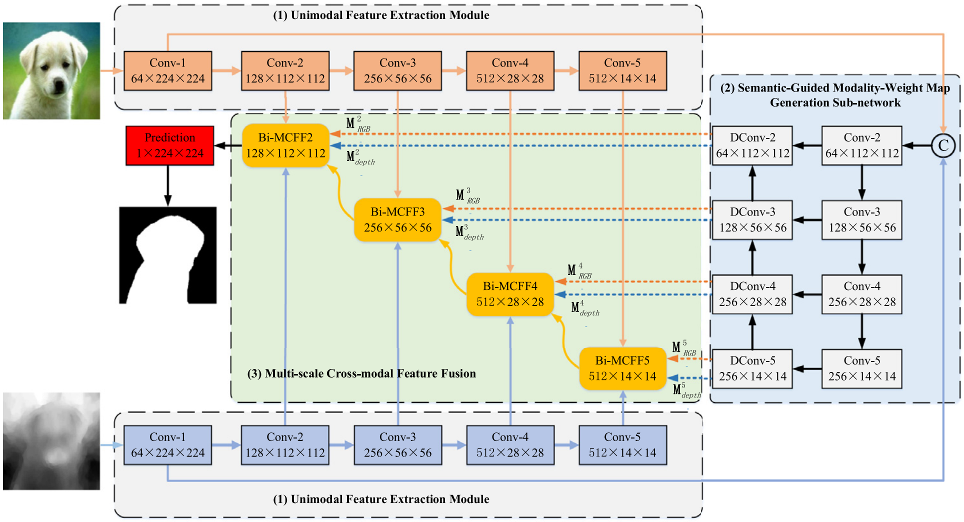

Fig. 4. Illustration of the proposed model. Given the input RGB-D images, the unimodal features are first extracted by employing the unimodal feature extraction module. Then, the modality-weight maps are generated by employing the proposed SG-MWMG sub-network. Finally, given the extracted unimodal features and the corresponding modality-weight maps, the proposed MCFF modules are employed to extract multi-level multi-scale cross-modal features for salient object detection.

In the decoder stage, the semantic information extracted in the encoder stage is employed to guide the generation of the modality-weight maps (i.e., $\mathbf{M}^i_{RGB}$ and $\mathbf{M}^i_{depth}$) for the unimodal features (i.e., $\mathbf{F}^i_{RGB}$ and $\mathbf{F}^i_{depth}$). Here, i = 2,3,4,5. Specifically, in the ith decoder stage, the modality-weight maps (i.e., $\mathbf{M}^i_{RGB}$ and $\mathbf{M}^i_{depth}$) for the ith level of unimodal RGB/depth features are generated as follows. First, the cross-level features $\tilde{\mathbf{H}}^{i+1}$ (if existed) from the $(i+1)$th decoder stage and the extracted features $\mathbf{H}^i$ from the ith encoder block are first concatenated. Then, the cross-level features $\tilde{\mathbf{H}}^i$ are generated by employing two stacked 3×3 convolutional layers on the concatenated features. Mathematically, this process can be expressed by

$$\tilde{\mathbf{H}}^i = \begin{cases} \mathrm{Conv}(\mathrm{Cat}(\mathrm{UP}(\tilde{\mathbf{H}}^{i+1}), \mathbf{H}^i), \theta), i = 2, 3, 4, \\ \mathrm{Conv}(\mathbf{H}^i, \theta), i = 5, \end{cases} \qquad (1)$$

where $\mathrm{UP}(*)$ denotes the up-sample operation by the bilinear interpolation algorithm. $\mathrm{Cat}(*)$ denotes the concatenation. $\mathrm{Conv}(*, \theta)$ refers to the convolutional layers with the parameter of θ.

Finally, two modality-weight maps $\mathbf{M}^i_{RGB}$ and $\mathbf{M}^i_{depth}$ in the ith level are obtained by performing a 1×1 convolutional layer on $\tilde{\mathbf{H}}^i$, which can be expressed by

$$\mathbf{M}^i_{RGB}, \mathbf{M}^i_{depth} = \mathrm{Conv}(\tilde{\mathbf{H}}^i, \gamma_i), i = 2, 3, 4, 5, \qquad (2)$$

where $\mathrm{Conv}(*, \gamma_i)$ denotes the 1×1 convolutional layer with the parameters γ_i at ith level. The modality-weight maps $\mathbf{M}^i_{RGB}$ and $\mathbf{M}^i_{depth}$ indicate the qualities of local regions in the input RGB/Depth images, respectively. Specifically, higher values in the modality-weight maps mean that the corresponding regions in RGB/depth images are more likely to be informative (high-quality) for the saliency prediction, and vice versa.

Some existing works also tried to address this issue. In [21], based on the observation that the high-quality depth images tend to have well-defined closed boundaries and show clear double peaks in their depth distribution, a depth depurator unit (DDU) was presented to determine whether a depth image should be used for saliency detection. However, the DDU is not end-to-end trainable. Moreover, only partial regions in some depth images may be non-informative and the rest of regions may still provide useful geometrical cues for saliency prediction, which will also be discarded by using DDU. In [8], based on the channel-wise attention mechanism, an attention-aware cross-modal cross-level fusion (ACCF) block was proposed to select the discriminative unimodal features. However, in the ACCF block, the features from a feature map of an input image will be wholly preserved (boosted) or discarded (suppressed) just accordingly to their global cues extracted by a global average pool. As a result, some features from a non-informative local region may be mistakenly preserved (or boosted) if their corresponding global activation is high, while some features from an informative local region may also be mistakenly discarded (suppressed) during the feature selection if their corresponding global activation is low.

Unlike these models, the proposed SG-MWMG sub-network locally determines the informative regions as well as the non-informative regions within each input RGB-D image guided by their corresponding semantic information. In the subsequent saliency detection, the features in those local informative regions are seen as discriminative ones and will be adaptively selected, while the features in those local non-informative regions will be seen as non-discriminative ones and will be discarded.

3.3. Bi-directional multi-scale cross-modal feature fusion (Bi-MCFF) module

Given the modality-weight maps generated by the proposed SG-MWMG sub-network, the next step is to effectively capture the cross-modal complementary information by selectively fusing those discriminative unimodal features for saliency detection. For that, a Bi-directional Multi-scale Cross-modal Feature Fusion (Bi-MCFF) module is presented. Unlike most of the existing RGB-D saliency detection models that fuse the unimodal features just at a single scale [26,31], the proposed Bi-MCFF module will extract the cross-modal features at multiple levels and scales from the input unimodal features. This may potentially increase the accuracy in identifying the salient objects.

The architecture of the proposed Bi-MCFF module is shown in Fig. 5. Given the unimodal features (i.e., $\mathbf{F}^i_{RGB}$ and $\mathbf{F}^i_{depth}$) at the ith level and the corresponding modality-weight maps (i.e., $\mathbf{M}^i_{RGB}$ and $\mathbf{M}^i_{depth}$), the multi-level multi-scale cross-modal features $\tilde{\mathbf{F}}^i_{RGB-D}$ are obtained by the following three steps.

First, selecting the discriminative unimodal features with the aid of those modality-weight maps generated by the proposed SG-MWMG sub-network. Specifically, the selected discriminative unimodal features (i.e., $\tilde{\mathbf{F}}^i_{RGB}$ and $\tilde{\mathbf{F}}^i_{depth}$) in the ith level are computed

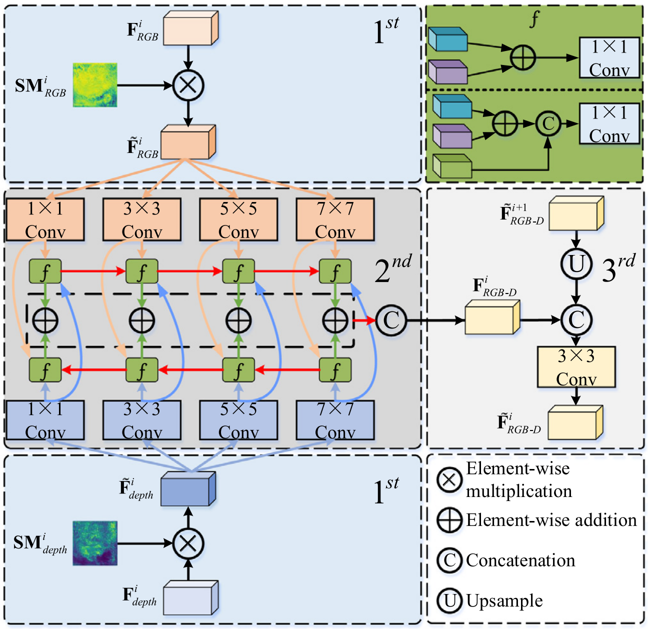

Fig. 5. Architecture of the proposed Bi-MCFF module. In Bi-MCFF, the unimodal features are first selected by using the modality-weight maps, and then the multi-scale unimodal features are extracted from the selected unimodal features. Finally, the multi-scale cross-modal features are extracted by fusing those multi-scale unimodal features at different scales.

by

$$\tilde{\mathbf{F}}^i_{RGB} = \mathbf{M}^i_{RGB} \odot \mathbf{F}^i_{RGB}, \\ \tilde{\mathbf{F}}^i_{depth} = \mathbf{M}^i_{depth} \odot \mathbf{F}^i_{depth}, \tag{3}$$

where $\odot$ denotes the element-wise multiplication. Here, i=2,3,4,5. $\mathbf{F}^1_{RGB}$ and $\mathbf{F}^1_{Depth}$ are not employed to capture the cross-modal information due to the fact that these features usually contain many more fine details as well as noises.

Second, capturing multi-scale cross-modal features $\mathbf{F}^i_{RGB-D}$ from the selected unimodal features (i.e., $\tilde{\mathbf{F}}^i_{RGB}$ and $\tilde{\mathbf{F}}^i_{depth}$). Specifically, as shown in Fig. 5, four levels of the unimodal features (denote by $\tilde{\mathbf{F}}^i_{m,l}$) are extracted by using four parallel convolutional layers with different kernel sizes (i.e., 1×1, 3×3, 5×5 and 7×7). Here $m \in \{RGB, depth\}$ denotes the unimodal RGB/depth features, and $l = 1, 2, 3, 4$ denotes different scales. Mathematically, they are obtained by

$$\tilde{\mathbf{F}}^i_{m,l} = \mathrm{Conv}\left(\tilde{\mathbf{F}}^i_m, \lambda^i_{m,l}\right), \tag{4}$$

where $\mathrm{Conv}\left(*, \lambda^i_{m,l}\right)$ denotes the convolutional layers with different kernel sizes and the corresponding parameters $\lambda^i_{m,l}$. $\tilde{\mathbf{F}}^i_m$ denotes the selected unimodal features at the ith ($i = 2, 3, 4, 5$) level.

There are cross-modal complementary information among those unimodal features and, meanwhile, cross-scale complementary information among those multi-scale features, due to the fact that the unimodal features with certain scales cannot cover the various sizes, shapes and positions of salient objects. To effectively capture the cross-scale and cross-modal complementary information, as shown by the red arrows in Fig. 5, a bi-directional cross-scale cross-modal feature fusion strategy is designed. Specifically, in the left-to-right direction, the cross-modal features at smaller scales are employed to facilitate the extraction of cross-modal features at larger scales. In contrast, in the right-to-left direction, the cross-modal features at larger scales are employed to facilitate the extraction of cross-modal features at smaller scales. Mathematically, this is expressed by

$$\overrightarrow{\mathbf{F}}^i_{RGB-D,l} = \begin{cases} \mathrm{Conv}(\tilde{\mathbf{F}}^i_{RGB,l} + \tilde{\mathbf{F}}^i_{depth,l}, \overrightarrow{\vartheta}^i_l), l = 1, \\ \mathrm{Conv}(\mathrm{Cat}(\overrightarrow{\mathbf{F}}^i_{RGB-D,l-1}, \tilde{\mathbf{F}}^i_{RGB,l} + \tilde{\mathbf{F}}^i_{depth,l}), \overrightarrow{\vartheta}^i_l), l = 2, 3, 4, \end{cases} \tag{5}$$

$$\overleftarrow{\mathbf{F}}^i_{RGB-D,l} = \begin{cases} \mathrm{Conv}(\tilde{\mathbf{F}}^i_{RGB,l} + \tilde{\mathbf{F}}^i_{depth,l}, \overleftarrow{\vartheta}^i_l), l = 4 \\ \mathrm{Conv}(\mathrm{Cat}(\overleftarrow{\mathbf{F}}^i_{RGB-D,l+1}, \tilde{\mathbf{F}}^i_{RGB,l} + \tilde{\mathbf{F}}^i_{depth,l}), \overleftarrow{\vartheta}^i_l), l = 3, 2, 1, \end{cases} \tag{6}$$

where $\overrightarrow{\mathbf{F}}^i_{RGB-D,l}$ and $\overleftarrow{\mathbf{F}}^i_{RGB-D,l}$ denote the extracted lth scale of cross-modal features at ith level. $\mathrm{Conv}(*, \overrightarrow{\vartheta}^i_l)$ and $\mathrm{Conv}(*, \overleftarrow{\vartheta}^i_l)$ denote the 1×1 convolutional layer with corresponding parameters $\overrightarrow{\vartheta}^i_l$ and $\overleftarrow{\vartheta}^i_l$, respectively. After that, the lth scale of cross-modal features (i.e., $\mathbf{F}^i_{RGB-D,l}$) are obtained by

$$\mathbf{F}^i_{RGB-D} = \mathrm{Cat}(\overrightarrow{\mathbf{F}}^i_{RGB-D,1} + \overleftarrow{\mathbf{F}}^i_{RGB-D,1}, \overrightarrow{\mathbf{F}}^i_{RGB-D,2} + \overleftarrow{\mathbf{F}}^i_{RGB-D,2}, \\ \overrightarrow{\mathbf{F}}^i_{RGB-D,3} + \overleftarrow{\mathbf{F}}^i_{RGB-D,3}, \overrightarrow{\mathbf{F}}^i_{RGB-D,4} + \overleftarrow{\mathbf{F}}^i_{RGB-D,4}). \tag{7}$$

In this way, the interactions of the features across different modalities and scales are exploited to effectively capture the cross-scale and cross-modal complementary information for RGB-D SOD.

Third, obtaining the multi-level multi-scale cross-modal features by fusing the multi-scale cross-modal features from two adjacent levels. This may provide more discriminative cross-modal features for the detection of salient objects. Specifically, as shown in the third part of Fig. 5, the multi-level multi-scale cross-modal features $\tilde{\mathbf{F}}^i_{RGB-D}$ of the ith level are obtained as follows. The (i+1)th level of features $\tilde{\mathbf{F}}^{i+1}_{RGB-D}$ (if existed) are upsampled, and then concatenated with the multi-scale cross-modal features $\mathbf{F}^i_{RGB-D}$ of the ith level. After that, a 3×3 convolutional layer is applied to extract the multi-level multi-scale cross-modal features. Mathematically, this process is expressed by

$$\tilde{\mathbf{F}}^i_{RGB-D} = \begin{cases} \mathrm{Conv}\left(\mathrm{Cat}\left(\mathrm{UP}\left(\tilde{\mathbf{F}}^{i+1}_{RGB-D}\right), \mathbf{F}^i_{RGB-D}\right), \theta^i\right), i = 2, 3, 4, \\ \mathrm{Conv}\left(\mathbf{F}^i_{RGB-D}, \theta^i\right), i = 5, \end{cases} \tag{8}$$

where $\mathrm{Conv}\left(*, \theta^i\right)$ is a 3×3 convolutional layer and θ^i is the corresponding parameters.

3.4. Loss function

Cross-entropy loss between the final saliency map $\mathbf{S}$ and the ground truth $\mathbf{Y}$ is employed to optimize the proposed model, which is computed by

$$\varsigma_1 = \mathbf{Y}\log(\mathbf{S}) + (1 - \mathbf{Y})\log(1 - \mathbf{S}). \tag{9}$$

Besides that, the intermediate supervisions as in Hao and Youfu [11] are also employed on the proposed Bi-MCFF modules to better capture the multi-scale cross-modal complementary information at different levels. Mathematically, the loss is expressed by

$$\varsigma_2 = \sum_{i=2}^{5} (\mathbf{Y}_i \log(\mathbf{S}_i) + (1 - \mathbf{Y}_i)\log(1 - \mathbf{S}_i)) \tag{10}$$

where $\mathbf{S}_i$ is the saliency map deduced by the output features of Bi-MCFF module in the ith level through a 1×1 convolution layer with the Sigmoid function. $\mathbf{Y}_i$ is the corresponding ground truth at ith level, which is sampled from $\mathbf{Y}$ as the size of $\mathbf{S}_i$.

However, the deduced saliency map $\mathbf{S}$ may have some blurring boundaries if only the cross-entropy loss is employed for training [11]. Therefore, a boundary-preserved loss is further employed to boost the details of the salient object boundaries. Specifically, two boundary maps are first deduced by employing the Sobel operator on the saliency map **S** and the corresponding ground truth **Y**. Then, the mean absolute error (*MAE*) between two boundary maps are computed to supervise the generation of salient object boundaries, which is expressed by

$$\varsigma_3 = |\,\mathrm{Sobel}(\mathbf{Y}) - \mathrm{Sobel}(\mathbf{S})|_1, \tag{11}$$

where $|*|_1$ denotes the l_1-norm of a matrix [32]. Therefore, the total loss ς for training the proposed model is expressed by

$$\varsigma = \varsigma_1 + \varsigma_2 + \varsigma_3. \tag{12}$$

4. Experiments

4.1. Datasets

The effectiveness of the proposed model is evaluated on four public datasets: NJU2000 [33], NLPR [34], STEREO [35] and SIP [21].

NJU2000 collects 1985 RGB-D image pairs from the internet, 3D movies and photographs captured by a Fuji W3 stereo camera. NLPR contains 1000 RGB-D image pairs under different scenarios captured by Kinect. STEREO consists of 797 RGB-D images pairs. SIP is a new dataset, which consists of 1000 accurately annotated high-resolution RGB-D image pairs.

For fair comparisons, we use the same training and testing set as in Fan et al. [21], Piao et al. [36] and Zhang et al. [12]. Concretely, 1485 RGB-D image pairs from the NJU2K dataset and 700 RGB-D image pairs from the NLPR dataset are sampled as our training set. The remaining images in the NJU2K and NLPR datasets and the whole datasets of STEREO and SIP are used for testing.

4.2. Evaluation metrics

Four evaluation metrics (i.e., F-measure (F_β), PR curve, mean absolute error (*MAE*) [37], S-measure (S_λ)) [38] and E-measure (E_γ) [39] are adopted to comprehensively evaluate various methods. For *MAE*, lower values are desirable. and for others higher values are desirable.

F-measure (F_β) evaluates the overall performance of a salient object detection model, which is a harmonic mean of *Precision* and *Recall*, and is expressed by:

$$F_\beta = \frac{(1+\beta^2)\times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{13}$$

where β^2 is set to 0.3, as suggested in Houwen et al. [33]. *Precision* and *Recall* are computed by comparing the ground truths and the binarized saliency maps under different thresholds.

MAE denotes the average absolute difference between the saliency map **S** and the ground truth **Y**, which is expressed by

$$MAE = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|\mathbf{S}(x,y) - \mathbf{Y}(x,y)|, \tag{14}$$

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

S-measure (S_λ) evaluates the spatial structure similarities between the saliency map **S** and the ground truth **Y**, which is formulated by

$$S_\lambda = \alpha * S_o + (1-\alpha) * S_r, \tag{15}$$

where $\alpha \in [0, 1]$ is the balance parameter and is set to 0.5 as default. More details are seen in Fan et al. [38].

E-measure (E_γ) [39] combines local pixel values with the image-level mean value in one term, thus jointly capturing image-level statistics and local pixel matching information. It is computed by:asure (E_γ) [39] combines local pixel values with the image-level mean value in one term, thus jointly capturing image-level statistics and local pixel matching information. It is computed by:

Table 1
Quantitative results of basic analysis.

Methods	*MAE*	F_β	S_λ	$E\gamma$
Baseline	0.0529	0.8696	0.8930	0.9098
+SG-MWMG	0.0439	0.8812	0.9061	0.9234
+Bi-MCFF	0.0453	0.8768	0.9040	0.9188
+SG-MWMG + Bi-MCFF	**0.0382**	**0.8952**	**0.9114**	**0.9347**

$$E_\gamma = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}\phi_{FM}(x,y), \tag{16}$$

where W and H are the width and height of saliency maps. $\phi_{FM}(*)$ is the enhanced alignment matrix whose details are in DengPing et al. [39].

4.3. Implementation

The proposed model is implemented with Pytorch toolbox and trained on an NVIDIA 1080Ti GPU. The loss function is minimized by using mini-batch Stochastic Gradient Descent (SGD) with the batch size of 4 and the momentum of 0.95. The learning rate is initialized as 0.002 and decreased by a factor of 0.8 for every 20 epochs. The input images are resized as 224 × 224 in the training process.

4.4. Ablation experiments and analyses

In this section, the ablation analysis of each component of our proposed model is performed on NJU2000 to investigate their validities and contributions.

4.4.1. Basic analysis

We first investigate the impact of each component of our proposed model. Specifically, the ‘Baseline’ model denotes the one that has removed the SG-MWMG sub-network and Bi-MCFF module from our proposed RGB-D SOD model. As shown in Table 1, both the SG-MWMG sub-network and Bi-MCFF module (i.e., ‘Baseline + SG-MWMG’ and ‘Baseline + Bi-MCFF’) can improve the performance of RGB-D SOD. Furthermore, with the collaboration of SG-MWMG sub-network and Bi-MCFF module, ‘Baseline + SG-MWMG + Bi-MCFF’ obtains the best performance.

4.4.2. SG-MWMG sub-network

To investigate the impact of the SG-MWMG sub-network, several versions of our proposed method (i.e., RGB_only, Depth_only, M_No_Sel, M_Input, M_Level1, M_Level2, MAP2_For_ALL and M_Level_ALL, for short, respectively) are provided for comparison. For RGB_only and Depth_only, only the unimodal RGB and depth images are employed for salient object detection, respectively. In M_No_Sel, the proposed SG-MWMG sub-network is removed from the proposed model, which means that the unimodal features are directly fed into the fusion module. In M_Input, instead of using the unimodal (RGB/depth) features from the first level as in M_Level1 (i.e., the proposed model), the input RGB and depth images are directly concatenated as the four-channel inputs of the SG-MWMG sub-network. Similarly, in ‘M_level2’, the unimodal RGB and depth features at the second level are concatenated as the inputs of our proposed SG-MWMG. In M_Level_ALL, the channels of the unimodal (RGB/depth) features from all of the levels (i.e., level1-level5) are first reduced to 64. Then, all of these reduced features are resized and concatenated as the inputs of our proposed SG-MWMG. The unimodal features from the third, fourth

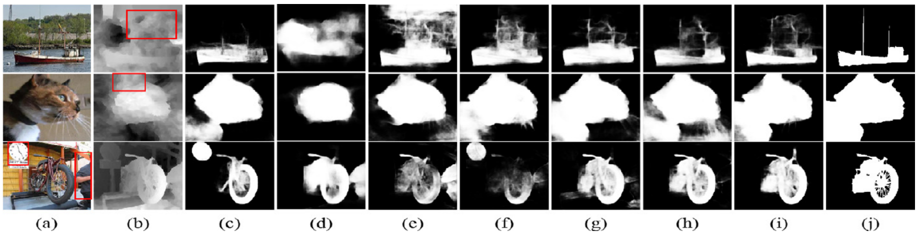

Fig. 6. Visual results of different selection modules. (a) Depth images; (b) RGB images; (c)–(i) Saliency maps deduced by RGB_only, Depth_only, M_No_SEL, M_input, M_Level2, M_Level_ALL and M_Level1; (j) Ground truth. As shown in the regions marked by the red boxes, the depth images in the first two rows contain some low-quality regions and the RGB image in the third row also contains some disturbing information. In those cases, as shown in (c) (d) and (e), the saliency maps deduced by fused cross-modal features may be worse than those by the unimodal features. Meanwhile, as shown in (f)–(i), the saliency maps generated by modules with selection are closer to the ground truths. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Quantitative results of different selection modules.

Methods	*MAE*	F_β	S_λ	$E\gamma$
Depth_only	0.0729	0.8106	0.8635	0.8694
RGB_only	0.0562	0.8513	0.8958	0.8989
M_No_SEL	0.0453	0.8768	0.9040	0.9188
M_Input	0.0402	0.8901	0.9102	0.9310
M_Level1	**0.0382**	**0.8952**	0.9114	**0.9347**
M_Level2	0.0386	0.8923	**0.9130**	0.9336
M_Level_ALL	0.0405	0.8893	0.9087	0.9301
MAP2_For_ALL	0.0431	0.8863	0.9081	0.9261

Table 3
Quantitative results of different selection models.

Methods	*MAE*	F_β	S_λ	$E\gamma$
SP_Sel	0.0397	0.8911	0.9105	0.9313
Ch_Sel	0.0428	0.8816	0.9078	0.9250
DMAP_only	0.0436	0.8822	0.9080	0.9241
RMAP_only	0.0468	0.8743	0.9044	0.9173
T_SG-MWMG	0.0404	0.8891	0.9094	0.9300
SG-MWMG	**0.0382**	**0.8952**	**0.9114**	**0.9347**

or fifth levels are not compared because these features mainly contain high-level semantic information. MAP2_For_ALL employs $\mathbf{M}^2_{RGB}$ and $\mathbf{M}^2_{depth}$ to select the discriminative unimodal RGB and depth features at all of the levels.

Quantitative evaluation The quantitative results of different models are shown in Table 2, which indicates that the performance of salient object detection may generally be boosted by employing the cross-modal complementary information (i.e., RGB_only and Depth_only obtain worse results than other models). However, as shown in Fig. 6, in some special cases, the saliency results obtained by using the fused cross-modal features may be worse than those obtained by using unimodal features. For example, the saliency maps deduced by RGB_only in the first two rows and the saliency map deduced by Depth_only in the third row are more accurate than those deduced by M_No_Sel. Meanwhile, the models with selection (i.e., M_Input, M_Level1, M_Level2 and M_Level_ALL) may obtain better results in those cases shown in Fig. 6.

The quantitative results shown in Table 2 also indicate that the models with selection (i.e., M_Input, M_Level1, M_Level2 and M_Level_ALL) outperform the models without selection (i.e., M_No_Sel) by a large margin. This owes to that those disturbing information contained in the low-quality images is suppressed by the selection of unimodal features. Therefore, more discriminative cross-modal features are extracted by fusing the selected discriminative unimodal features from those models with selection than by simply fusing all of the input unimodal features without selection. Moreover, M_Level1 (i.e., the proposed model) performs the best among these models.

Furthermore, directly employing $\mathbf{M}^2_{RGB}$ and $\mathbf{M}^2_{depth}$ for the unimodal RGB and depth features at all of the levels (i.e., MAP2_For_ALL) obtains suboptimal results. This may be due to the fact that the features at different levels facilitate salient object detection from different aspects, e.g., the features at high levels are mainly to locate the salient objects, while the features at low levels are mainly to obtain the boundaries of the salient objects. Fig. 7 visualizes some modality-weight maps for RGB and depth images at different levels. It can be seen that, for the unimodal features at high levels, the modality-weight maps generated by our SG-MWMG sub-network tend to focus on those informative regions of RGB and depth images due to the fact that extracting more information from those informative regions of RGB and depth images may help to locate the salient objects. While, for the features at low levels, the modality-weight maps generated by our SG-MWMG sub-network tend to focus on those salient regions due to the fact that paying more attention to the salient objects may help to capture more boundary information for accurately segmentation.

Comparisons with other selection modules In addition, some other feature selection modules are also employed for comparison with our proposed SG-MWMG sub-network. First, two existing attention mechanism based unimodal feature selection modules are employed, i.e., SP_Sel [40] and Ch_Sel [8]. Here, SP_Sel adopts the same structure as in Fig. 8(a) and the Ch_Sel adopts the same structure as in Fig. 8(b). Then, instead of simultaneously using modality-weight maps for unimodal RGB and depth features, DMAP_only and RMAP_only only employ one of the modality-weight maps for unimodal RGB and depth features, respectively. Besides, another version of the SG-MWMG sub-network is designed (i.e., T_SG-MWMG, for short). Instead of taking the concatenated unimodal features from the first level as the inputs, T_SG-MWMG employs two independent sub-networks to generate the modality-weight maps for the unimodal features, respectively, where each of the sub-networks takes the unimodal RGB/depth features from the first level as the inputs. For better visualization, the four levels of modality-weight maps generated by these selection modules are first resized to the same size. Then the average of those four resized modality-weight maps are computed and displayed in Fig. 9. The quantitative results of these modules are also shown in Table 3.

As shown in Table 3, the proposed SG-MWMG sub-network achieves better performance than others. This may owe to the following facts. SP_Sel may assign larger weights to the salient regions and smaller weights to the non-salient regions. Accordingly, as shown in Fig. 9(g) and (h), the modality-weight maps generated by SP_Sel mainly focus on the salient regions in the input images. As a result, as shown in the first and the second rows of Fig. 9, SP_Sel may obtain good results in some simple scene. However, as shown in the third row of Fig. 9, SP_Sel may be powerless to detect salient objects in some complex scenes. Similarly, Ch_Sel

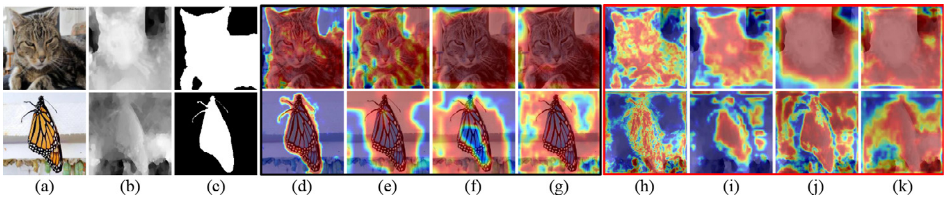

Fig. 7. Visualization of modality-weight maps at different levels. (a) RGB images. (b) Depth images. (c) Ground truth. (d)–(g) Modality-weight maps for RGB images at the first, second, third and fourth levels, respectively. (h)–(k) Modality-weight maps for depth images at the first, second, third and fourth levels, respectively.

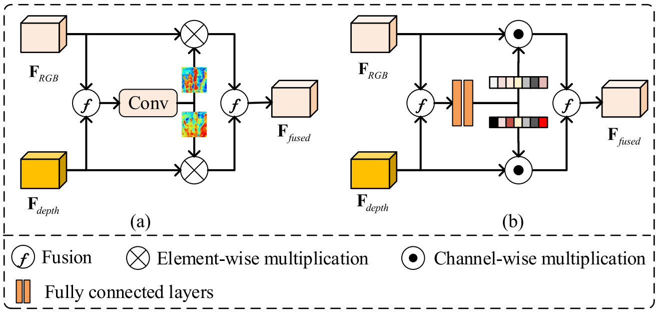

Fig. 8. Architectures of existing feature selection modules. (a) Spatial-wise feature selection; (b) Channel-wise feature selection..

also obtains suboptimal results. Furthermore, only using one of the modality-weight maps for selecting unimodal RGB or depth features (i.e., DMAP_only or RMAP_only) also obtains suboptimal results, due to the fact that, in practice, both the RGB images and the depth images may have low qualities.

Differently, SG-MWMG and T_SG-MWMG can be seen as an improved version of existing attention mechanism based unimodal feature selection modules, which pays more attention to those informative regions of the input images, regard less of salient or non-salient ones. As a result, as shown in Fig. 9, all of the features within these informative regions will benefit the final saliency detection with the guidance of semantic information. Thus, better results may be obtained for those complex scenes. Meanwhile, in T_SG-MWMG, only the unimodal features from one of the modalities are employed to generate the modality-weight maps, while in SG-MWMG sub-network, the cross-modal features are jointly employed to generate the modality-weight maps, which may better determine the informative regions within the RGB-D images.

4.4.3. Bi-MCFF module

In order to demonstrate the validity of the proposed Bi-MCFF module, another three versions of our proposed Bi-MCFF module (i.e., No_MCFF, A_MCFF, B_MCFF, and MCFF, for short, respectively) are designed for comparison. In No_MCFF, the proposed Bi-MCFF modules are replaced by two stacked residual blocks [30] to fuse the unimodal features. It should be noted that No_MCFF contains more parameters than the proposed module. A_MCFF follows the same structure as in Fig. 3(a), where the selected unimodal features are first concatenated and then fused by employing a 3 × 3 convolution layer on the concatenated features. After that, four parallel convolutional layers with different kernel sizes are employed to extract the multi-scale cross-modal features. B_MCFF takes the same structure as in Fig. 3(b), where the multi-scale unimodal features are first extracted independently from the corresponding unimodal features by using four parallel convolutional layers with different kernel sizes. Then, all scales of unimodal RGB and depth features are concatenated and fused by employing a 3 × 3 convolution layer. In MCFF, the bi-directional structure are removed from our proposed Bi-MCFF module. Furthermore, the results of using different fusion directions in Bi-MCFF module are also evaluated. Specifically, in ‘Left-MCFF’ and ‘Right-MCFF’, both the two paths fuse the multi-scale features in the same direction, i.e., left-to-right direction and right-to-left direction, respectively, rather than different directions as in Bi-MCFF module.

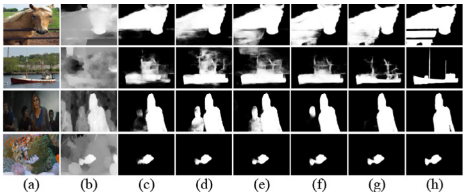

Fig. 10. Illustration of some saliency results deduced by different versions of the proposed Bi-MCFF module. (a) RGB images; (b) Depth images; (c)–(g) Saliency maps deduced by No_MCFF, A_MCFF, B_MCFF, MCFF and Bi-MCFF; (h) Ground truth.

Fig. 10 illustrates some salient objects of different sizes. As shown in Fig. 10, No_MCFF only detects some parts of those salient objects, while other modules identify more accurate or even complete ones in those scenes. This owes to that, compared with No_MCFF, other modules (i.e., A_MCFF, B_MCFF, MCFF and Bi-MCFF) extract multi-scale features from input RGB-D images, which are more robust to the sizes of salient objects. Moreover,

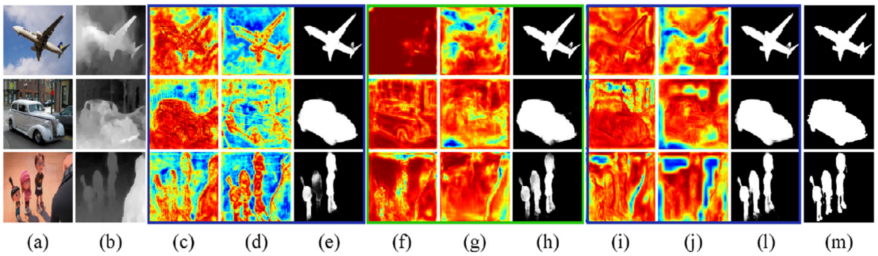

Fig. 9. Visual results of modality-weight maps generated by different selection modules. (a) RGB images; (b) Depth images; (c) and (d) Modality-weight maps for RGB and depth images generated by SP_Sel; (e) Saliency maps generated by SP_Sel; (f) and (g) Modality-weight maps for RGB and depth images generated by T_SG-MWMG; (h) Saliency maps generated by T_SG-MWMG; (i)-(j) Modality-weight maps for RGB and depth images generated by SG-MWMG; (l) Saliency maps generated by SG-MWMG; (m) Ground truth. Regions with higher values (i.e., marked by the red color) in modality-weight maps are more likely to be informative for saliency detection. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

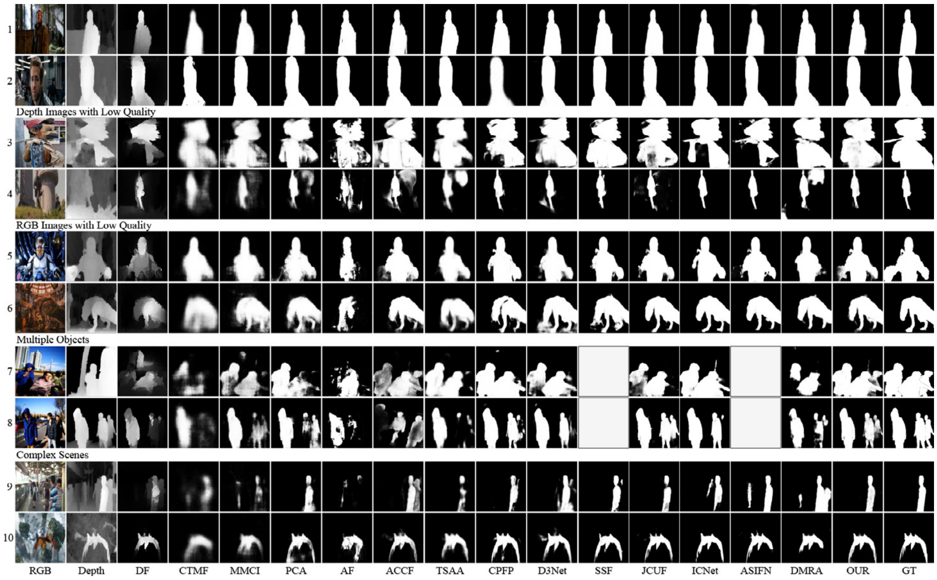

Fig. 11. Visualization of saliency maps generated by different models.

Table 4
Quantitative results of different versions of the proposed Bi-MCFF module.

Methods	*MAE*	F_β	S_λ	$E\gamma$
No_MCFF	0.0439	0.8812	0.9061	0.9234
A_MCFF	0.0425	0.8841	0.9072	0.9272
B_MCFF	0.0422	0.8854	0.9086	0.9274
MCFF	0.0394	0.8907	0.9112	0.9321
Left-MCFF	0.0391	0.8901	**0.9123**	0.9347
Right-MCFF	0.0385	0.8920	0.9108	0.9345
Bi-MCFF	**0.0382**	**0.8952**	0.9114	**0.9347**

compared with A_MCFF, B_MCFF and MCFF, the proposed Bi-MCFF module achieves more complete and accurate saliency maps. This may result from the fact that more cross-scale and cross-modal complementary information can be obtained by employing the bi-directional structure.

As shown in Table 4, the quantitative results reveal that the proposed Bi-MCFF module obtains the best performance. For A_MCFF, some multi-scale information within one of the modalities may not appear in the fused features, since the unimodal features are fused firstly. Similarly, for B_MCFF, all of the multi-scale unimodal features are directly fused to extract the multi-scale cross-modal features. As a result, one scale of the unimodal features may also be drowned by the features from other scales. While, in the MCFF and Bi-MCFF module, the unimodal features are first selected, and then the multi-scale cross-modal features are separately extracted by fusing the unimodal RGB/depth features from the same scales. In this way, the discriminative information contained in one scale of unimodal features may not be drowned by those non-discriminative unimodal features from other scales or modalities. As a result, the multi-scale cross-modal complementary features within RGB-D image are better exploited.

The results of ‘Left-MCFF’ and ‘Right-MCFF’ indicate that fusing multi-scale features only from one direction cannot fully exploit the cross-scale complementary information. Furthermore, by virtue of Bi-MCFF module, the cross-scale complementary information is well exploited in two separate directions, where the interactions of the features across modalities and scales are effectively exploited. As a result, the proposed Bi-MCFF module obtains better saliency detection results.

4.5. Comparisons to state-of-the-art models

The proposed model and some existing state-of-the-art RGB-D salient object detection models are evaluated on four benchmark

Table 5
Quantitative results of different models. The best three results are marked by the colors of Red, Blue and Green.

Datasets	NJU2000				NLPR				STEREO				SIP			
	MAE	F_β	S_λ	$E\gamma$	*MAE*	F_β	S_λ	$E\gamma$	*MAE*	F_β	S_λ	$E\gamma$	*MAE*	F_β	S_λ	$E\gamma$
DF [41]	0.141	0.649	0.762	0.696	0.084	0.664	0.801	0.754	0.140	0.616	0.757	0.691	0.185	0.464	0.652	0.564
MMCI [27]	0.078	0.793	0.858	0.851	0.059	0.736	0.855	0.841	0.067	0.812	0.872	0.873	0.086	0.770	0.832	0.844
CTMF [26]	0.084	0.778	0.849	0.846	0.056	0.740	0.859	0.840	0.086	0.758	0.848	0.841	0.139	0.607	0.715	0.704
PCA [31]	0.059	0.839	0.876	0.895	0.043	0.802	0.873	0.887	0.063	0.818	0.874	0.887	0.070	0.814	0.842	0.878
ACCF [8]	0.055	0.861	0.898	0.907	0.033	0.860	0.909	0.923	0.048	0.866	0.894	0.911	0.070	0.817	0.866	0.878
TSAA [11]	0.060	0.841	0.879	0.895	0.041	0.819	0.886	0.901	0.059	0.827	0.871	0.893	0.075	0.803	0.834	0.870
AF [28]	0.099	0.765	0.773	0.826	0.058	0.755	0.799	0.850	0.075	0.806	0.824	0.872	0.117	0.701	0.720	0.792
CPFP [25]	0.053	0.850	0.895	0.910	0.035	0.840	0.888	0.917	0.051	0.841	0.879	0.912	0.063	0.820	0.850	0.893
D3Net [21]	0.051	0.860	0.895	0.912	0.033	0.852	0.905	0.923	0.048	0.844	0.890	0.908	0.062	0.832	0.864	0.882
JCUF [42]	0.041	0.881	0.901	0.926	0.030	0.885	0.900	0.932	0.045	0.870	0.895	0.924	0.056	0.854	0.873	0.904
ICNet [43]	0.052	0.869	0.894	0.913	0.029	0.884	0.926	0.939	0.044	0.869	0.902	0.925	0.068	0.834	0.853	0.029
ASIFN [44]	0.047	0.881	0.901	0.926	0.030	0.885	0.900	0.932	0.045	0.870	0.895	0.924	–	–	–	–
DMRA [36]	0.050	0.873	0.885	0.919	0.031	0.864	0.898	0.939	0.048	0.867	0.885	0.930	0.085	0.819	0.805	0.843
SSF [12]	0.043	0.884	0.897	0.927	0.026	0.881	0.913	0.946	0.045	0.877	0.892	0.928	–	–	–	–
OUR	0.038	0.895	0.911	0.934	0.025	0.891	0.926	0.948	0.041	0.874	0.904	0.930	0.047	0.877	0.888	0.922

datasets: NJU2000 [33], NLPR [34], STEREO [35] and SIP [21]. The eight existing state-of-the-art models include DF [41], MMCI [27], CTMF [26], PCA [31], ACCF [8], TSAA [11], AF [28], CPFP [25], D3Net [21], JCUF [42], ICNet [43], ASIFN [44], DMRA [36] and SSF [12]. For fair comparisons, the saliency maps deduced by these existing models are provided by their authors and are tested by our evaluating code.

Quantitative results of different models are illustrated in Table 5. It can be seen that, on NJU2000, NLPR and SIP datasets, the proposed method outperforms others in terms of all the evaluation metrics. While, on the STEREO dataset, our proposed model achieves the best results in terms of *MAE*, E_γ and E_γ, and obtains competitive results with respect to F_β. Visualization results under different scenarios are illustrated in Fig. 11. As shown in the first two rows of Fig. 11, for the images under simple scenes, most of these methods succeed to detect the salient objects. However, when one of the input images are low-quality (e.g., the images shown in the 3rd – 6th rows), most of the existing methods fail to detect the whole objects or lose some fine details, while our method can well detect the whole salient regions. This may owe to the proposed SG-MWMG subnetwork. By virtue of SG-MWMG, those discriminative unimodal features may be adaptively preserved and those non-discriminative unimodal features may be also discarded. Accordingly, more discriminative cross-modal features are extracted from the selected unimodal features for saliency detection. Meanwhile, when there are multiple salient objects in the input RGB-D images (e.g., the images shown in the 7th and 8th rows), most of the existing models also fail to identify the salient objects, while the proposed model can successfully detect all of the salient objects. This may owe to the Bi-MCFF module in our proposed method. By using the proposed Bi-MCFF module, multi-level multi-scale cross-modal features are well captured from the input RGB-D images. The incorporation of the proposed SG-MWMG and Bi-MCFF modules helps the proposed model to obtain more accurate saliency maps than other models under complex scenes (e.g., the images shown in the 9th and 10th rows of Fig. 11).

5. Conclusion and future work

In this paper, a novel RGB-D salient object detection model has been presented, where the qualities of the input images are taken into consideration during saliency prediction. For that, an SG-MWMG sub-network is specially designed in our model. Exploring SG-MWMG, the informative (high-quality) and non-informative (low-quality) regions within the input images can be adaptively determined with the guidance of semantic information extracted from the RGB-D images. The importance of the unimodal RGB/depth features at each fusion stage can also be weighted, which in turn helps to highlight informative features but suppress non-informative features during the subsequent saliency prediction. This significantly boosts the performance of salient object detection, especially when one of the input images has low quality. Moreover, an Bi-MCFF module is designed to extract cross-scale and cross-modal complementary information by using the designed bi-directional structure. By virtue of Bi-MCFF, our proposed model can still work well in the scenes that contain multiple salient objects of different sizes. The incorporation of SG-MWMG and MCFF further helps the proposed model to achieve desirable saliency detection results under complex scenes. Experimental results on several benchmarks validated the validity of the proposed model.

Our proposed model may be further improved from the following aspects in the future. First, using an independent sub-network (i.e., SG-MWMG sub-network) for generating weight maps introduces lots of parameters, which may be simplified in the future. Secondly, the simplest multi-modal feature fusion strategy (i.e., element-wise addition) is employed, which ignores the interactions between the unimodal RGB and depth features. More feasible multi-modal feature fusion strategies, which considers the interactions between the unimodal RGB and depth features, may be designed in the future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant no. 61773301. It is also supported by the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University.

References

[1] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[2] B.Y. Lei, E.-L. Tan, S. Chen, D. Ni, T. Wang, Saliency-driven image classification method based on histogram mining and image score, Pattern Recognit. 48 (2015) 2567–2580.
[3] Q. Cai, H. Liu, Y. Qian, S. Zhou, X. Duan, Y. Yang, Saliency-guided level set model for automatic object segmentation, Pattern Recognit. 93 (2019) 147–163.
[4] W. Hua, D. Cheng, Y. Junchi, T. Dacheng, Asymmetric cross-guided attention network for actor and action video segmentation from natural language query, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3938–3947.
[5] Y. Kong, M. Feng, X. Li, H. Lu, X. Liu, B. Yin, Spatial context-aware network for salient object detection, Pattern Recognit. 114 (2021) 107867.
[6] Q. Zhang, Y. Shi, X. Zhang, Attention and boundary guided salient object detection, Pattern Recognit. 107 (2020) 107484–107496.
[7] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, L. Shao, RGB-D salient object detection: a survey, Comput. Vis. Media 7 (2021) 1–33.
[8] C. Hao, L. Youfu, S. Dan, Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2018, pp. 6821–6826.
[9] L. Duan, W. Ma, Y. Qiao, Z. Cai, J. Miao, Q. Ye, Cocnn: RGB-D deep fusion for stereoscopic salient object detection, Pattern Recognit. 104 (2020) 107329.
[10] Q. Chen, K. Fu, Z. Liu, G. Chen, H. Du, B. Qiu, L. Shao, Ef-net: a novel enhancement and fusion network for RGB-D saliency detection, Pattern Recognit. 112 (2021) 107740.
[11] C. Hao, L. Youfu, Three-stream attention-aware network for RGB-D salient object detection, IEEE Trans. Image Process. 28 (2019) 2825–2835.
[12] M. Zhang, W. Ren, Y. Piao, Z. Rong, H. Lu, Select, supplement and focus for RGB-D saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3472–3481.
[13] L. Yi, H. Jungong, Z. Qiang, W. Long, Salient object detection via two-stage graphs, IEEE Trans. Circuits Syst. Video Technol. 29 (2019) 1023–1037.
[14] Y. Liu, Z. Qiang, H. Jungong, W. Long, Salient object detection employing robust sparse representation and local consistency, Image Vis. Comput. 69 (2017) 155–167.
[15] Z. Qiang, L. Yi, Z. Siyang, H. Jungong, Salient object detection based on super-pixel clustering and unified low-rank representation, Comput. Vis. Image Underst. 161 (2017) 51–64.
[16] D. Cheng, Y. Xu, N. Feiping, T. Dapeng, Saliency detection via a multiple self--weighted graph-based manifold ranking, IEEE Trans. Multimed. 22 (2019) 885–896.
[17] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 4 (2017) 640–651.
[18] L. Yi, H. Jungong, Z. Qiang, S. Caifeng, Deep salient object detection with contextual information guidance, IEEE Trans. Image Process. 29 (2019) 360–374.
[19] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, P.H.S. Torr, Deeply supervised salient object detection with short connections, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 815–828.
[20] C. Runmin, L. Jianjun, F. Huazhu, H. Junhui, H. Qingming, K. Sam, Going from RGB to RGB-D saliency: a depth-guided transformation model, IEEE Trans. Cybern. 50 (8) (2019) 3627–3639.
[21] D. Fan, Z. Lin, J.-X. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, M. Cheng, Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks, IEEE Trans. Neural Netw. Learn. Syst. 32 (2021) 2075–2089.
[22] J. Zhang, D. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, N. Barnes, Uncertainty inspired RGB-D saliency detection, IEEE Trans Pattern Anal. Mach. Intell. 17 (2021) 13–19.

[23] Z. Chunbiao, C. Xing, H. Kan, H.L. Thomas, L. Ge, PDNet: prior-model guided depth-enhanced network for salient object detection, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2018, pp. 199–204.

[24] F. Liang, L. Duan, W. Ma, Y. Qiao, J. Miao, Q. Ye, Context-aware network for RGB-D salient object detection, Pattern Recognit. 111 (2021) 107630.

[25] Z. JiaXing, C. Yang, F. DengPing, C. Mingming, L. Xuanyi, Z. Le, 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.

[26] H. Junwei, C. Hao, L. Nian, Y. Chenggang, L. Xuelong, CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion, IEEE Trans. Cybern. 48 (2018) 3171–3183.

[27] C. Hao, L. Youfu, S. Dan, Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. 86 (2019) 376–385.

[28] N. Wang, X. Gong, Adaptive fusion for RGB-D salient object detection, IEEE Access 7 (2019) 55277–55284.

[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, 2015, pp. 1–5.

[30] H. Kaiming, Z. Xiangyu, R. Shaoqing, S. Jian, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 770–778.

[31] C. Hao, L. Youfu, 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.

[32] N. Feiping, W. Hua, D. Cheng, G. Xinbo, L. Xuelong, H. Heng, New l2, 1-norm relaxation of multi-way graph cut for clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 1962–1968.

[33] P. Houwen, L. Bing, X. Weihua, H. Weiming, J. Rongrong, RGB-D salient object detection: a benchmark and algorithms, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 92–109.

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

[35] N. Yuzhen, G. Yujie, L. Xueqing, L. Feng, Leveraging stereopsis for saliency analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 454–461.

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

[37] A. Borji, M. Cheng, H. Jiang, J. Li, Salient object detection: a benchmark, IEEE Trans. Image Process. 24 (12) (2015) 5706–5722.

[38] D. Fan, M. Cheng, Y. Liu, T. Li, 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.

[39] F. DengPing, G. Cheng, C. Yang, R. Bo, C. MingMing, B. Ali, Enhanced-alignment measure for binary foreground map evaluation, in: International Joint Conference on Artificial Intelligence, 2018, pp. 698–704.

[40] Q. Zhang, T. Xiao, N. Huang, D. Zhang, J. Han, Revisiting feature fusion for RGB-T salient object detection, IEEE Trans. Circuits Syst. Video Technol. 31 (2020) 804–1818.

[41] Q. Liangqiong, H. Shengfeng, Z. Jiawei, T. Jiandong, T. Yandong, Y. Qingxiong, RGB-D salient object detection via deep fusion, IEEE Trans. Image Process. 26 (2017) 2274–2285.

[42] H. Nianchang, L. Yi, Z. Qiang, H. Jungong, Joint cross-modal and unimodal features for RGB-D salient object detection, IEEE Trans. Multimed. 23 (2020) 2428–2441.

[43] G. Li, Z. Liu, H. Ling, ICNet: Information conversion network for RGB-D based salient object detection, IEEE Trans. Image Process. 29 (2020) 4873–4884.

[44] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, Q. Huang, ASIF-Net: attention steered interweave fusion network for RGB-D salient object detection, IEEE Trans. Cybern. 51 (1) (2020) 88–100.

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.

Yongjiang Luo received the B.S. degree in automatic control, the M.S. degree and Ph.D. in circuit and system from Xidian University, China, in 2001, 2004, and 2011, respectively. He was a Visiting Scholar with University of California, Merced, USA. He is currently an associate professor with the School of Electronic Engineering, Xidian University, China. His current research interests include wideband signal processing and intelligent information processing.

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 and 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.