Image and Vision Computing 95 (2020) 103888

Contents lists available at ScienceDirect

Image and Vision Computing

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

Attention-guided RGBD saliency detection using appearance information ☆

Xiaofei Zhou[a], Gongyang Li[b,c], Chen Gong[d], Zhi Liu[b,c,*], Jiyong Zhang[a]

[a] Institute of Information and Control, Hangzhou Dianzi University, Hangzhou 310018, China
[b] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China
[c] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
[d] The Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

ARTICLE INFO

Article history:
Received 5 January 2020
Accepted 26 January 2020
Available online 1 February 2020

Keywords:
RGBD
Saliency
Bottom-up
Top-down
Attention
Appearance

ABSTRACT

Most of the deep convolutional neural networks (CNNs) based RGBD saliency models either regard the RGB and depth cues as the same status or trust the depth information excessively. However, they ignore that the low-quality depth map is an interference factor and the multi-level deep features that originated from RGB images contain abundant appearance information. Therefore, we propose a novel RGBD saliency model, where the attention-guided bottom-up and top-down modules are powerfully combined by using multi-level deep RGB features, to utilize the deep RGB and depth features in a sufficient and appropriate way. Concretely, a two-stream structure based bottom-up module is first constructed to dig and fuse the RGB and depth information, yielding the fused deep feature. Besides, the module embeds the depth cue based attention maps to guide the indication of salient objects. Then, considering the abundant appearance information, a top-down module is deployed to perform coarse-to-fine saliency inference, where the fused deep feature is progressively integrated with appearance information. Similarly, the attention map is also inserted into this module for locating salient objects. Extensive experiments are performed on five public RGBD datasets and the corresponding experimental results firmly demonstrate the effectiveness and superiority of our model when compared with the state-of-the-art RGBD saliency models.

1. Introduction

Saliency detection aims to highlight the most visually discriminative objects in some scenes and has received increasing concern in recent years. It has a wide variety of applications in many fields such as image segmentation [1,2], quality assessment [3], action recognition [4] and person re-identification [5]. Correspondingly, there are many effective efforts in this field, in particular, the traditional machine learning based models [6-10] together with the convolutional neural networks (CNNs) based models [11-15] significantly promote the performance of saliency detection. However, these existing saliency models mainly focus on RGB images or videos. Meanwhile, with the recent development of RGBD sensing technologies such as Microsoft Kinect, mobile phone camera and RealSense, depth information, which provides complementary information about objects' spatial structure and layout, can be obtained conveniently. On the basis of this, the booming research topic, *i.e.* RGBD saliency detection, which employs RGB image and depth map together to pop-out the most attractive objects in RGBD scenes, draws more and more attention due to its practical application value in many visual analysis tasks, such as object detection [16], visual tracking [17], and image retargeting [18].

For RGBD saliency models, most of the early efforts [19-30] mainly design hand-crafted features to construct various models, such as contrast-based models [23], background-based model [25], cellular automata based model [27], minimum barrier distance based model [29] and multi-scale fusion based model [28]. However, the performance of the aforementioned efforts often declines significantly when dealing with some complex scenes such as cluttered background, low contrast and heterogeneous objects. This can be mainly attributed to the lack of high-level semantic information, because most of the hand-crafted features are low-level features, which cannot provide effective representations for salient objects and background regions. Fortunately, in recent years, deep learning

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

* Corresponding author.
E-mail addresses: zxforchid@outlook.com (X. Zhou), ligongyang@shu.edu.cn (G. Li), chen.gong@njust.edu.cn (C. Gong), liuzhisjtu@163.com (Z. Liu), jzhang@hdu.edu.cn (J. Zhang).

https://doi.org/10.1016/j.imavis.2020.103888
0262-8856/

techniques especially the CNNs, which is often used for extracting high-level semantic features, are widely deployed to RGBD saliency models [31-42], such as complementarity-aware fusion network [35], three-stream model [41] and fluid pyramid integration model [42], which further boost the RGBD saliency detection's performance.

Among these RGBD saliency models, there are two major models. The first one is the single-stream model, such as the efforts [28,38,36,42], which inserts the depth information into the core branch, *i.e.* RGB branch and achieves a comparable performance when compared with the state-of-art models. The other major model adopts two-stream structure [20,23,21,26,34,35,40,39], which tries to aggregate the RGB and depth information via the multi-level and multi-modal fusion architecture, and also achieves an encouraging performance. However, a further promotion for RGBD saliency detection is still imperative, especially when dealing with some challenging scenes including small object, unclear depth, complex objects and so on. Meanwhile, through a thorough study, we find that the aforementioned efforts treat the RGB and depth information equally or rely on the depth information heavily. Unfortunately, as we know, the inherent attributes of RGB and depth information are considerably different, and they may lead to some incompatible faults when combined by using simple operations (including concatenation, summation and multiplication), some complicated methods as well as symmetric information interaction between the paired modality. Besides, due to the technique limitation of sensors, depth maps usually contain amount of noise and their quality is lower than RGB images. Furthermore, for the multi-level deep features generated from RGB branch, namely the appearance information, it is also beneficial for the depiction of salient objects, where the top layers contain rich semantic information and the bottom layers indicate spatial details. Thus, according to this, we can say that the usage of RGB and depth information is worthy of serious consideration.

Motivated by the aforementioned argumentation, this paper proposes a novel RGBD saliency model to utilize the RGB and depth information sufficiently and appropriately. To achieve this goal, we devise three key components. First, bottom-up and top-down strategies have been successfully applied in saliency models [43-45], and the U-shape architecture based or fully convolutional networks (FCN) based saliency models [11,12,15], which are mainly an encoder-decoder architecture, have also achieved an encouraging performance. Thus, inspired by this, our entire network also adopts a U-shape or encoder-decoder architecture, which contains bottom-up and top-down modules. Meanwhile, considering the successful employment of two-stream structure in RGBD saliency models, we design a two-stream structure based bottom-up module to perform feature extraction and fusion for the RGB and depth modalities. Besides, as the aforementioned discussion about multi-level deep features obtained from RGB branch in the bottom-up module, we pay more attention to RGB information, namely the abundant appearance information, via the top-down module, in which the fused deep feature generated by bottom-up module gradually integrates appearance information from top layers to bottom layers. Following this way, the bottom-up module and the top-down module are further combined using multi-level deep RGB features. Lastly, enlightened by the effective utilization of attention maps [46,14], we also embed the attention maps, which can be obtained from depth cue by using the contrast-enhanced method [42], into the entire network to guide the differentiation of salient objects and background regions. In this way, we can not only ease the negative impact of some low-quality depth maps but also further promote the sufficient utilization of the depth information.

Therefore, our model shown in Fig. 1 first puts the RGB image and the depth map into the bottom-up module, where we can extract deep RGB and depth features and generate the fused deep feature. Then, the top-down module is deployed to progressively integrate the fused deep feature with the multi-level deep RGB features, namely the appearance information. During this process, the fine-grained saliency prediction is performed in a top-down way. Finally, we can get the high-quality saliency map, *i.e.* the output of the top-down module. Overall, the main contribution of this proposed model can be presented as follows:

1. We propose a novel RGBD saliency model, in which the attention-guided two-stream structure based bottom-up module and top-down module are strongly linked using appearance information, to utilize the RGB and depth information sufficiently and appropriately.
2. To sufficiently utilize the depth information and effectively ease the adverse impact of low-quality depth maps, we devise attention maps to both modules for guiding the discrimination of salient objects and background regions.
3. To adequately use the RGB information, we pay more attention to the appearance information, *i.e.* the multi-level deep RGB features, which is progressively integrated with the fused deep feature in a top-down manner.
4. Comparing with the state-of-the-art RGBD saliency models on five public RGBD datasets, our model shows distinguished performance, and this clearly demonstrates the effectiveness of the proposed RGBD saliency model.

The remaining of this paper is organized as follows. The related works about RGBD saliency detection are reviewed in Section 2. Section 3 gives a detailed description of the proposed RGBD saliency model. The experimental results and the corresponding analyses are discussed in Section 4. Finally, we draw a conclusion for this paper in Section 5.

2. Related works

In recent years, numerous efforts including the hand-crafted based models [19-30] and the deep learning based models [31-42] have devoted to perform RGBD saliency detection, and achieve encouraging performance. Among these models, the exploration in the complementarity of RGB and depth information is the key issue in the research of RGBD saliency model. Usually, these models can be categorized into two classes, namely the single-stream RGBD saliency model and the two-stream RGBD saliency model. Therefore, in this section, we mainly give a brief review for the two RGBD saliency models.

2.1. Single-stream RGBD saliency model

Many efforts have been devoted to the single-stream RGBD saliency models. For example, in [28], various hand-crafted features are aggregated via a random forest regressor, and the final saliency map is the summation of all saliency estimations in all scales. In [31], various saliency feature vectors such as local/global contrast and spatial/background prior are further processed to generate the saliency probability for each superpixel, and then Laplacian propagation is employed to generate the spatial consistent saliency maps. In [32], the top-down and bottom-up cues are utilized via a deep learning based architecture, in which the superpixel-based hand-crafted depth features are combined with the deep RGB features to obtain saliency scores by using fully-connected layers. In [38], a master network is designed for extracting deep RGB features and a sub-network is utilized to extract deep depth features, which are incorporated into the master network. In [36], Liu et al. proposed a single stream recurrent convolution neural network (SSRCNN), in which the RGB and depth maps are concatenated into a four-channels input. In [42], the proposed fluid pyramid network tries to integrate the multi-level deep features to generate saliency maps for RGBD images. In

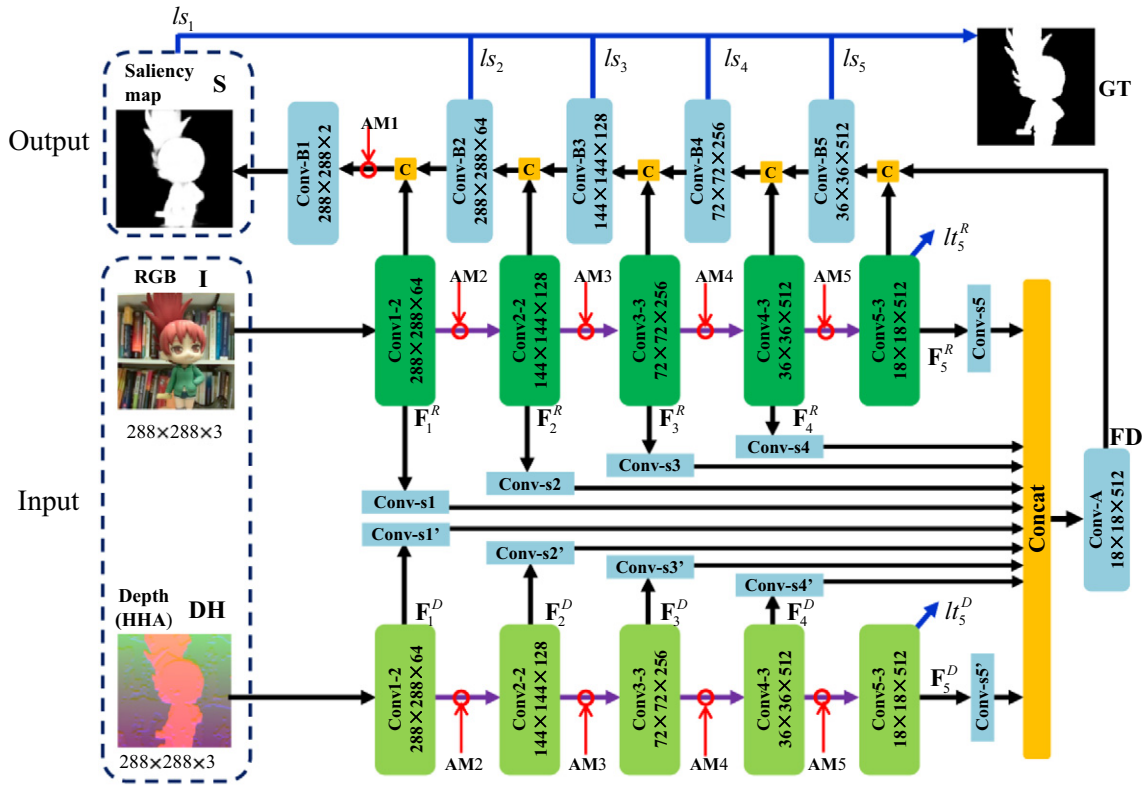

Fig. 1. Illustration of the proposed RGBD saliency model: under the guidance of attention maps $\{\text{AM}i\}_{i=1}^{5}$, a two-stream structure based bottom-up module first generates multi-level and multi-modal deep features $\{\mathbf{F}_i^R\}_{i=1}^{L}$ and $\{\mathbf{F}_i^D\}_{i=1}^{L}$ ($L = 5$), which are further processed by convolutional layers (Conv-*si* and Conv-*si*′, $i = 1, \cdots, 5$) and then aggregated into a fused deep feature **FD** using the concatenation (Concat) and convolution (Conv-A). Successively, through the top-down module, the fused deep feature **FD** is progressively integrated with the appearance information, *i.e.* the multi-level deep RGB features $\{\mathbf{F}_i^R\}_{i=1}^{L}$, via concatenation layers (C) and convolutional blocks (Conv-B*i*, ($i = 1, \cdots, 5$)). Following this way, we can obtain a high-quality saliency map **S** for each input RGB image **I** and depth map **DH**.

comparison with the aforementioned single-stream RGBD saliency models, our model is unique. For our entire network, we devise a encoder-decoder-like architecture, where the two-stream structure based bottom-up module (*i.e.* encoder) and the top-down module (*i.e.* decoder) are trained jointly. During the whole process, our model gives more concern to appearance information and introduces depth cue based attention maps, which further strengthen the combination of both modules and guide the salient objects' location, respectively. Following this way, we can acquire a high-quality saliency map for each RGBD image.

2.2. Two-stream RGBD saliency model

For the two-stream RGBD saliency models, some prior works are constructed by using hand-crafted features. For example, in [20], depth contrast, color contrast and spatial bias, which can be obtained using color and depth values, are combined together to generate visual saliency maps via multiplication operation. In [23], four kinds of contrast are first computed using color, luminance, texture and depth cues, and then they are fused into the final saliency map, in which the fusion method allocates different weights to different contrast maps according to their compactness values. In [21], a multi-stage saliency model is proposed, where the feature contrast based map, region grouping based map together with location and scale based map are all computed using color and depth information simultaneously. In [26], the color and depth values based compactness saliency map is combined with the foreground saliency map, which is generated based on depth cue and multiple contrast, via weighted-sum approach. In [27], an evaluation strategy based RGBD saliency model is proposed, where the color channel based saliency map and the depth channel based saliency map are first fused and then refined using cellular automata.

With the wide application of convolutional neural networks (CNNs), the CNNs-based RGBD saliency models also perform very well. For instance, in [34], except the exploration in the complementarity of RGB and depth information, the RGB branch and the depth branch are integrated using a combination layer, *i.e.* a fully connected layer. In [35], the proposed model aims to explore the complementarity of RGB and depth cues via a fusion network, which effectively integrates the cross-modal and cross-level deep features. Successively, in [40], Chen et al. proposed another two-stream fusion model, which not only diversifies the paths in RGB and depth branches but also deploys interactions into many layers. In [39], the fusion module learns a switch map to adaptively integrate the estimated saliency results from RGB and depth branches. Generally, the two-stream structure based bottom-up module in our model also inherits the merits of the aforementioned two-stream RGBD saliency models for feature mining and aggregation. Differently, our model also introduces attention maps for guiding the following process including feature extraction, fusion and salient objects' location. Furthermore, the bottom-up module is strongly combined with the top-down module by using appearance information, namely the multi-level deep RGB features, where the generated fused deep feature is progressively integrated with appearance information, yielding the high-quality saliency map.

3. The proposed method

In this part, we first give a brief introduction for the proposed RGBD saliency model in Section 3.1. Then, the bottom-up module

is detailed in Section 3.2. Next, we discuss the top-down module in Section 3.3. Lastly, the training and implementation details will be elaborated in Section 3.4.

3.1. Overall architecture

The proposed RGBD saliency model shown in Fig. 1 contains a two-stream structure based bottom-up module and top-down module, which are effectively combined using appearance information and are jointly trained in an end-to-end manner. Concretely, the entire network of our model is an encoder-decoder architecture, and the bottom-up module (*i.e.* encoder) is built based on the basic network VGG-16 model [47]. Firstly, the input of our model is the RGB image **I** and the depth map **DH**, which are all in the size of $W \times H$ (W means the width and H is the height). Notably, it is unsuitable for the original depth map **D** to directly use the pre-trained model of VGG-16, thus we encode the original depth map **D** into a three-channel HHA image **DH** using [48]. Then, the input RGB image **I** and depth map **DH** are passed into the bottom-up module to perform feature extraction and fusion, yielding the fused deep feature **FD**. During this process, we can also obtain multi-level and multi-modal deep features $\{\mathbf{F}_i^R\}_{i=1}^L$ and $\{\mathbf{F}_i^D\}_{i=1}^L$ ($L = 5$), *i.e.* the deep RGB and depth features in different convolutional blocks. Among these deep features, $\{\mathbf{F}_i^R\}_{i=1}^L$ contains abundant appearance information, which can not only indicate low-level spatial details but also present high-level semantic cues. Successively, the fused deep feature **FD** is progressively aggregated with appearance information $\{\mathbf{F}_i^R\}_{i=1}^L$ via the top-down module (*i.e.* decoder). Following this way, we can obtain a high-quality saliency map **S** for each RGBD image.

3.2. Bottom-up module

To sufficiently dig and fuse the RGB and depth information, and motivated by the existing two-stream RGBD saliency models [39,35,34], we design the two-stream structure based bottom-up module to extract and fuse the deep RGB and depth features. Meanwhile, considering the negative impact of low-quality depth maps and inspired by the contribution of attention maps [46,14], we also introduce attention maps to the bottom-up module for guiding the deep feature mining. In the following, a detailed description is presented for the bottom-up module.

Firstly, the two-stream structure based bottom-up module contains two sibling branches shown in Fig. 1 including a RGB branch and a depth branch, which are devised to extract deep RGB features and deep depth features, respectively. Concretely, both branches are constructed based on VGG-16, and here, according to most of CNNs-based saliency models, we also remove the last three fully connected layers and the last pooling layer. Therefore, for each branch of the bottom-up module, it consists of thirteen convolutional layers and four max pooling layers, corresponding to five convolutional blocks shown in Fig. 1. Besides, the two branches with the same structure are devised to share the same weights in all the corresponding convolutional layers, but they are all equipped with batch normalization (BN) layers [49], which can indicate different domain knowledge (RGB and depth) from the viewpoint of statistics, between the convolutional layers and the ReLU layers. Following this way, we can acquire deep features with five-level resolutions, which are corresponding to Conv1-2 (in which the number of channels are set to 64), Conv2-2 (128 channels), Conv3-3 (256 channels), Conv4-3 (512 channels) and Conv5-3 (512 channels), in each branch. Here, for simplicity, Fig. 1 only shows the aforementioned five convolutional layers. Meanwhile, we denote the generated multi-level and multi-modal deep RGB and depth features are denoted as $\{\mathbf{F}_i^R\}_{i=1}^L$ and $\{\mathbf{F}_i^D\}_{i=1}^L$ ($L = 5$), which are all in the size of $\frac{W}{2^{i-1}}, \frac{H}{2^{i-1}}\rfloor$ and refer to the RGB branch and the depth branch, respectively.

Secondly, the bottom-up module also contains several other layers. Concretely, after obtaining the multi-level and multi-modal deep features $\{\mathbf{F}_i^R\}_{i=1}^L$ and $\{\mathbf{F}_i^D\}_{i=1}^L$ ($L = 5$), we choose to fuse them to obtain the fused deep feature. However, these deep features including $\{\mathbf{F}_i^R\}_{i=1}^L$ and $\{\mathbf{F}_i^D\}_{i=1}^L$ are with different resolutions, thus we employ five pairs of convolutional layers, namely Conv-*si* and Conv-*si*′ ($i = 1, \cdots, 5$), to downsample these deep feature maps and make these deep features with the same resolution as $\{\mathbf{F}_i^R\}^L$ or $\{\mathbf{F}_i^D\}^L$. Here, for each pair of convolutional layers Conv-*si* and Conv-*si*′, the kernel size is set to $2^{5-i} \times 2^{5-i}$, the stride size is equal to 2^{5-i} and the number of channels is set to 64. Successively, we deploy a concatenation layer (Concat) and a convolutional block (Conv-A) to effectively integrate the multi-level and multi-modal deep features, yielding the fused deep feature **FD**. The corresponding process can be defined as:

$$\mathbf{FD} = h^A\left(\left[\left[h_i^{SR}\left(\mathbf{F}_i^R\right)\right]_{i=1}^5, \left[h_i^{SD}\left(\mathbf{F}_i^D\right)\right]_{i=1}^5\right]\right), \tag{1}$$

where $[.]$ means concatenation, and h^A denotes the convolutional block Conv-A, which contains a convolutional layer (kernel size $= 3 \times 3$, stride size $= 1$, channels $= 512$), a BN layer and a ReLU layer. Besides, h_i^{SR} and h_i^{SD} refer to the pair of convolutional layers Conv-*si* and Conv-*si*′, respectively.

Lastly, the bottom-up module also introduces attention maps to guide the feature extraction, which will be beneficial for discriminating salient objects and background regions. Specifically, the attention maps $\{\mathrm{AM}i\}_{i=2}^5$ with four resolutions $\lfloor\frac{W}{2^{i-1}}, \frac{H}{2^{i-1}}\rfloor$ ($i = 1, \cdots, 5$) are first generated based on depth maps by using the contrast-enhanced method [42]. Then, we embed these attention maps into later four levels of RGB and depth branches. Notably, for simplicity, we call this operation the "guidance item", which contains the purple line (pooling and copy), the red circle (embed) and the red line (copy), as shown in Figs. 1 and 2(a). Thus, the input of *ith* guidance item, *i.e.* the multi-level and multi-modal deep feature $\mathbf{F}_{i-1}^*$ ($*$ denotes R or D), is processed in this way:

$$\mathbf{F}_{i-1}^{e*} = P\left(\mathbf{F}_{i-1}^*\right) + Conv\left(P\left(\mathbf{F}_{i-1}^*\right) \odot \mathbf{AM}i\right), \tag{2}$$

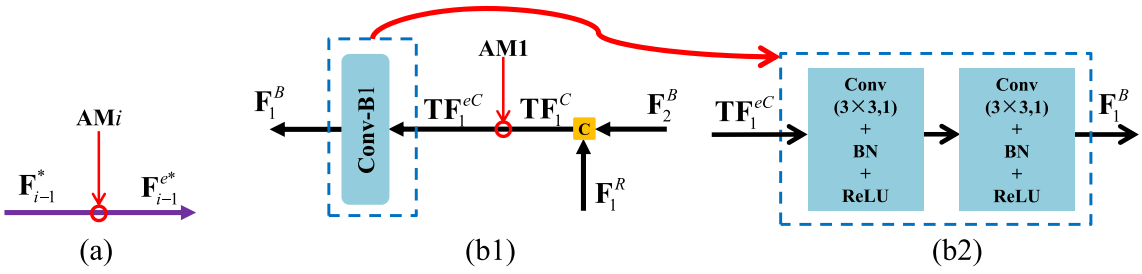

Fig. 2. (a): Illustration of the guidance operation of $\{\mathrm{AM}i\}_{i=2}^5$, (b1): Illustration of the guidance procedure of AM1, and (b2): the detailed configuration of convolutional block Conv-B1, which consists of convolutional layers (kernel size $= 3 \times 3$, stride $= 1$), batch normalization layers (BN) and ReLU layers. Noted, in (b1), "C" means concatenation.

where $\mathbf{F}_{i-1}^{e*}$ is the output of the *ith* guidance item in both branches, and it can be regarded as the enhanced deep feature including $\mathbf{F}_{i-1}^{eR}$ and $\mathbf{F}_{i-1}^{eD}$, which are also the input of following convolutional block Conv*i*. Besides, P is the pooling layer between two convolutional blocks, $\odot$ means pixel-wise multiplication and *Conv* is the convolutional layer (kernel size = 3×3, stride size =1). Notably, in this part, only $\{\text{AM}i\}_{i=2}^{5}$ participate in the guidance operation.

3.3. Top-down module

Through the bottom-up module, we can not only obtain the fused deep feature **FD** but also acquire the abundant appearance information, *i.e.* the multi-level deep RGB features $\{\mathbf{F}_i^R\}_{i=1}^{L}$ ($L = 5$). Besides, as we know, the multi-level deep RGB features provide powerful high-level semantic cues and low-level spatial details, which are useful for differentiating salient objects and background regions. Thus, we employ a decoder-like top-down module shown in Fig. 1 to gradually integrate the fused deep feature **FD** with the appearance information $\{\mathbf{F}_i^R\}_{i=1}^{L}$ ($L = 5$) and perform saliency estimation in a coarse-to-fine way. During this process, we give the RGB information more concern, which further boosts the combination of bottom-up and top-down modules. Meanwhile, similar as the bottom-up module, *i.e.* considering the negative influence of low-quality depth maps, we also deploy attention map into this part for guiding the location of salient objects.

According to Fig. 1, the top-down module contains five convolutional blocks Conv-B*i* ($i = 1,2,3,4,5$), and here, we divide it into two parts: one part includes the last four convolutional blocks Conv-B*i* ($i = 2,3,4,5$), which consists of convolutional layers (kernel size = 3 × 3, stride = 1), batch normalization (BN) layers, ReLU layers, deconvolutional layer (kernel size = 3 × 3, stride = 2) and dropout layer (ratio=0.5) [50], and the other part is the first convolutional block Conv-B*i* ($i = 1$). Concretely, firstly, for the *ith* convolutional blocks Conv-B*i* ($i = 2,3,4,5$) shown in Fig. 3, the detailed procedure can be defined as follows:

$$\mathbf{F}_i^B = h_i^B\left(\left[\mathbf{F}_{i+1}^B, \mathbf{F}_i^R\right]\right) \ (2 \leqslant i \leqslant 5), \tag{3}$$

where $\mathbf{F}_i^B$ denotes the output of each convolutional block Conv-B*i*, which is represented by h_i^B. Notably, for the last convolutional block Conv-B5, we initialize the $\mathbf{F}_{i+1}^B$ ($i = 5$) by **FD**, namely the input of the top-down module. During this process, we can see that the appearance information $\mathbf{F}_i^R$ together with $\mathbf{F}_{i+1}^B$ are passed into the following convolutional block Conv-B*i*, as shown in Fig. 3. Thus, following this way, we can obtain the intermediate result $\mathbf{F}_2^B$ from convolutional block Conv-B2, which only consists of two sub-convolutional blocks and is slightly different from those three sub-convolutional blocks based Conv-B*i* ($i = 3,4,5$). Similarly, the convolutional block Conv-B1 also contains two sub-convolutional blocks.

Secondly, according to Fig. 2 (b1), the intermediate result $\mathbf{F}_2^B$ first concatenates with deep RGB feature $\mathbf{F}_1^R$, yielding TF_1^C, which will be further enhanced by using the attention map **AM**1 ($W \times H$). Then, the enhanced feature TF_1^{eC} is passed into the first convolutional block Conv-B1 shown in Fig. 2 (b2), yielding the final deep feature $\mathbf{F}_1^B$. Thus, this guidance procedure can be formulated as:

$$\begin{aligned}
\mathbf{TF}_1^C &= \left[\mathbf{F}_2^B, \mathbf{F}_1^R\right] \\
\mathbf{TF}_1^{eC} &= \mathbf{TF}_1^C + Conv\left(\mathbf{TF}_1^C \odot \mathbf{AM}1\right) \\
\mathbf{F}_1^B &= h_1^B\left(\mathbf{TF}_1^{eC}\right),
\end{aligned} \tag{4}$$

where h_1^B denotes the Conv-B1. Similar to the guidance operation in Eq. (2), *Conv* and $\odot$ also refer to convolution and pixel-wise multiplication, respectively. Notably, the guidance operation of **AM**1 is slightly different from **AM***i* ($i = 2,\cdots,5$) shown in Fig. 2, where there is no pooling operation in Eq. (4). Eventually, the final high-quality saliency map **S** shown in Fig. 1 can be obtained by performing softmax operation on $\mathbf{F}_1^B$.

3.4. Model learning and implementation

The entire network of our model is a U-shape or encoder-decoder-like architecture, which can be trained in an end-to-end manner, and the encoder part, namely the two-stream structure based bottom-up module, is constructed based on VGG-16 [47]. Besides, inspired by some deeply-supervised saliency models [51,52,36], we not only deploy the deep supervision to all convolutional blocks Conv-B*i* ($i = 1,2,3,4,5$) in the top-down module, but also appoint it to the last convolutional blocks of RGB and depth branches in bottom-up module. Furthermore, with the deployment of attention maps, the corresponding loss (denoted as attention loss) will also be added in our total loss. Here, except for the attention loss, all the aforementioned loss are marked in blue lines, as shown in Fig. 1. Meanwhile, our total loss $\mathcal{L}$ can be denoted as:

$$\mathcal{L} = lt_5^R + lt_5^D + l_a + \sum_{i=1}^{5} ls_i, \tag{5}$$

where ls_i and l_a refer to the convolutional block Conv-B*i* and the attention loss, respectively. lt_5^R and lt_5^D respectively denote the loss of the last convolutional blocks of RGB branch and depth branch. Notably, ls_i, lt_5^R and lt_5^D are all computed using the cross-entropy loss.

We implement the proposed RGBD saliency model by using the Caffe toolbox [53] on a PC, which is equipped with an i7-4790 K

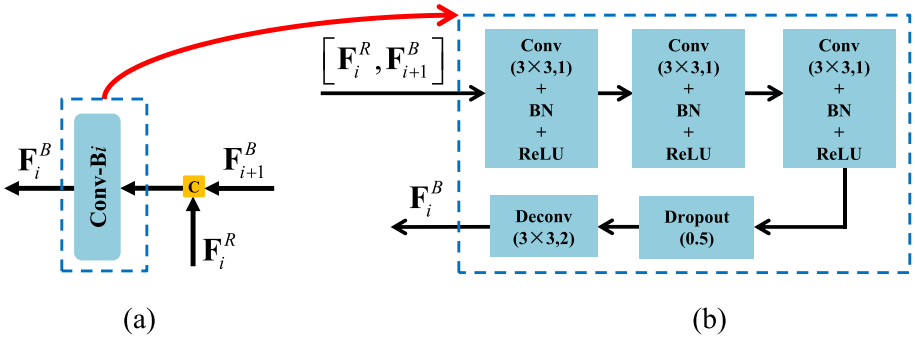

Fig. 3. Illustration of the convolutional block Conv-B*i* ($i = 2,\cdots,5$). (a): the thumbnail of Conv-B*i*, (b): the detailed configuration of Conv-B*i*, which consists of convolutional layers (kernel size = 3 × 3, stride = 1), batch normalization (BN) layers, ReLU layers, deconvolutional layer (kernel size = 3 × 3, stride = 2) and dropout layer (ratio=0.5). Noted, in (a), "C" in yellow box denotes concatenation.

CPU, a 32 GB memory RAM and a NVIDIA Titan XP GPU (with 12 GB memory). Specifically, we first initialize the parameters of the RGB and depth branches of the bottom-up module with the VGG-16 model, and the remaining parameters are initialized by using the "xavier" method [54]. Then, the stochastic gradient descent (SGD) is employed to minimize the total loss shown in Eq. (5), in which we set dropout ratio, batch size, mini-batch size, momentums and weight decay as 0.5, 1, 8, 0.9 and 0.0001, respectively. Besides, the learning rate is initialized to 10^{-7}, and it is consecutively divided by 10 after each 12,500 iterations. Here, we set the total training iterations as 25,000 for convergence. Furthermore, to train the proposed RGBD saliency model, we employ the same training set as CPFP [42], which randomly selects 1400 samples from the NJU2 K [22] and 650 samples from the NLPR [21]. In addition, we also adopt augmentation operations, in which we perform rotation on the original image with angles $90^{\circ}, 180^{\circ}$, and 270° and also adopt mirror reflection for the original image. In this way, we can totally obtain 10,250 training triplets including the RGB images, the depth maps and the ground truths.

4. Experimental results

This section first presents the RGBD datasets and evaluation metrics in Section 4.1. Then, in Section 4.2, we will compare our model with the state-of-the-art RGBD saliency models from the perspective of quantitative and qualitative views. Lastly, the detailed ablation studies will be shown in Section 4.3.

4.1. Datasets and evaluation metrics

4.1.1. Datasets

To comprehensively validate our model, we conduct extensive comparisons on five public RGBD datasets including NJU2K [22], NLPR [21], STEREO [19], LFSD [55] and DES [20]. In the following, we will give a brief introduction for these datasets: **NJU2K** contains 2003 image pairs including RGB images and depth maps, which are obtained from daily life, Internet and 3-D movies. Notably, 1400 image pairs in the training set are collected from NJU2K, and for testing, 485 image pairs in NJU2 K are taken out to construct a testing set named "**NJU2K-TE**". **NLPR** is built by using Microsoft Kinect and contains 1000 image pairs, in which some images possess multiple salient objects. To train our model, 650 image pairs in NLPR are also selected to make up the training set. Similarly, we also collect 300 image pairs from NLPR to constitute a testing set denoted as "**NLPR-TE**". **STEREO** is also called SSB1000 and has 1000 image pairs, which are mainly collected from Internet images and 3D movies. Here, all images in STEREO are regarded as testing set. **LFSD** and **DES** contain 100 and 135 image pairs, respectively, and both of them are regarded as testing set. Notably, these five RGBD datasets are all provided with pixel-wise annotated ground truth.

4.1.2. Evaluation metrics

To quantitatively make a comparison for all RGBD saliency models, we employ four evaluation metrics including S-measure (S) [56], max F-measure (maxF), max E-measure (maxE) [57] and mean absolute error (MAE) to evaluate all models' performance.

F-measure is treated as a comprehensive evaluation, which is the weighted harmonic mean of precision and recall, and it is defined as:

$$F_{\beta} = \frac{(1+\beta^2)\, Precision \times Recall}{\beta^2 Precision + Recall}, \tag{6}$$

where β^2 is set to 0.3 as suggested in [10,58]. We can obtain the max F-measure by using different thresholds [0,255].

MAE presents a balanced comparison between ground truth **GT** and saliency map **S**, which is formulated as:

$$\mathrm{MAE} = \frac{1}{W * H} \sum_{i=1}^{W*H} \left| \mathbf{S}(i) - \mathbf{GT}(i) \right|, \tag{7}$$

where W and H mean the width and height of the saliency map, respectively. Noted, to compute MAE, *i.e.* Eq. (7), **S** is firstly scaled to [0,1] for all RGBD saliency models.

S-measure is a structural similarity measure, which simultaneously incorporates the region-aware (S_r) and the object-aware (S_o) to measure the structural similarity between saliency map and ground truth. According to [56], the corresponding definition is shown as:

$$S = \alpha * S_o + (1-\alpha) * S_r, \tag{8}$$

where α denotes the balance parameter and is set to 0.5.

E-measure means the enhanced-alignment measure, which considers the local details and global information simultaneously. According to [57], the E-measure is defined as:

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

where $f(\cdot)$ can be defined as a convex function such as the quadratic form and $\circ$ is the Hadamard product. Besides, alignment matrix ξ is built on the bias matrices φ_{GT} and φ_{FM}, which are regarded as the distance between each pixel value of ground truth and its mean value respectively.

4.2. Comparison with the state-of-the-arts

To validate the performance of our model, we compare the proposed RGBD saliency model with totally 13 state-of-the-art RGBD saliency models including CDCP [30], ACSD [22], LBE [25], DCMC [26], SE [27], MDSF [28], DF [31], AFNet [39], CTMF [34], PCF [35], MMCI [40], TANet [41] and CPFP [42] on the aforementioned five public RGBD datasets including NJU2K-TE, NLPR-TE, STEREO, LFSD and DES. Here, the former six models are the classical models (*i.e.* non-deep learning), and the later seven models are all CNNs-based models. Notably, the saliency maps of all the 13 state-of-the-art models are generated by running the source codes or provided by the authors, and they are scaled to the same resolution as original images. Next, we will show the quantitative and qualitative comparisons successively.

The quantitative comparison results on five public challenging RGBD datasets are shown in Table 1 including S-measure, max F-measure, max E-measure and MAE. It can be seen that our model consistently outperforms the 13 state-of-art RGBD saliency models on most evaluation items, which convincingly demonstrates the superiority and effectiveness of the proposed RGBD saliency model. Besides, we also present the computation cost of all RGBD saliency models shown in the 2*nd* row of Table 1. As the aforementioned description in Section 3.4, our model is running on PC with an i7-4790 K CPU, a 32 GB memory RAM and a NVIDIA Titan XP GPU (with 12 GB memory) and is implemented via the Caffe toolbox [53]. Noted, we perform the experiments on 288×288 images. Referring to Table 1, we can see that the average running time of each image processed by our model is 0.179 s, which is a comparable result when compared with the 13 state-of-the-art RGBD saliency models.

Fig. 4 shows the qualitative comparisons for our model and 13 state-of-the-art RGBD saliency models on several challenging examples, which exhibit many attributes such as heterogeneous objects,

Table 1
Quantitative results of different RGBD saliency models on five public RGBD datasets in terms of S-measure, max F-measure, max E-measure and MAE. Besides, the average running time (seconds) of these models is also provided here. Noted, the best results are marked in bold face, and ↑ & ↓ mean that the larger and smaller one is better, respectively.

	Metric	CDCP [30]	ACSD [22]	LBE [25]	DCMC [26]	SE [27]	MDSF [28]	DF [31]	AFNet [39]	CTMF [34]	PCF [35]	MMCI [40]	TANet [41]	CPFP [42]	**Ours**
	Time	>60.0	0.718	3.110	1.200	1.570	>60.0	10.36	0.030	0.630	0.060	0.050	0.070	0.170	0.179
NJUD	*S*↑	0.669	0.699	0.695	0.686	0.664	0.748	0.763	0.772	0.849	0.877	0.858	0.878	0.879	**0.893**
TE	*maxF* ↑	0.621	0.711	0.748	0.715	0.748	0.775	0.804	0.775	0.845	0.872	0.852	0.874	0.877	**0.891**
	maxE ↑	0.741	0.803	0.803	0.799	0.813	0.838	0.864	0.853	0.913	0.924	0.915	0.925	0.926	**0.930**
	MAE ↓	0.180	0.202	0.153	0.172	0.169	0.157	0.141	0.100	0.085	0.059	0.079	0.060	**0.053**	0.055
NLPR	*S*↑	0.727	0.673	0.762	0.724	0.756	0.805	0.802	0.799	0.860	0.874	0.856	0.886	0.888	**0.914**
TE	*maxF* ↑	0.645	0.607	0.745	0.648	0.713	0.793	0.778	0.771	0.825	0.841	0.815	0.863	0.867	**0.897**
	maxE ↑	0.820	0.780	0.855	0.793	0.847	0.885	0.880	0.879	0.929	0.925	0.913	0.941	0.932	**0.950**
	MAE ↓	0.112	0.179	0.081	0.117	0.091	0.095	0.085	0.058	0.056	0.044	0.059	0.041	0.036	**0.031**
STERE	*S*↑	0.713	0.692	0.660	0.731	0.708	0.728	0.757	0.825	0.848	0.875	0.873	0.871	0.879	**0.893**
	maxF ↑	0.664	0.669	0.633	0.740	0.755	0.719	0.757	0.823	0.831	0.860	0.863	0.861	0.874	**0.886**
	maxE ↑	0.786	0.806	0.787	0.819	0.846	0.809	0.847	0.887	0.912	0.925	0.927	0.923	0.925	**0.930**
	MAE ↓	0.149	0.200	0.250	0.148	0.143	0.176	0.141	0.075	0.086	0.064	0.068	0.060	**0.051**	0.053
LFSD	*S* ↑	0.717	0.727	0.729	0.753	0.692	0.700	0.791	0.738	0.796	0.794	0.787	0.801	0.828	**0.876**
	maxF ↑	0.703	0.763	0.722	0.817	0.786	0.783	0.817	0.744	0.791	0.779	0.771	0.796	0.826	**0.877**
	maxE ↑	0.786	0.829	0.797	0.856	0.832	0.826	0.865	0.815	0.865	0.827	0.839	0.847	0.872	**0.912**
	MAE ↓	0.167	0.195	0.214	0.155	0.174	0.190	0.138	0.133	0.119	0.112	0.132	0.111	0.088	**0.070**
DES	*S* ↑	0.709	0.728	0.703	0.707	0.741	0.741	0.752	0.770	0.863	0.842	0.848	0.858	0.872	**0.913**
	maxF ↑	0.631	0.756	0.788	0.666	0.741	0.746	0.766	0.728	0.844	0.804	0.822	0.827	0.846	**0.897**
	maxE ↑	0.811	0.850	0.890	0.773	0.856	0.851	0.870	0.881	0.932	0.893	0.928	0.910	0.923	**0.951**
	MAE ↓	0.115	0.169	0.208	0.111	0.090	0.122	0.093	0.068	0.055	0.049	0.065	0.046	0.038	**0.031**

cluttered background, low contrast between salient objects and background regions as well as unclear depth. Specifically, from 1*st* to 5*th* rows, the images present heterogeneous salient objects with cluttered background, especially the 1*st*(the woman in blue vest with bare arms), 2*nd*(the girl with green shirt in red hair) and 3*rd*(the orange box with colorful slogan) examples. Thus, it is a challenging task for all models to perform saliency detection on these scenes. Fortunately, their depth maps are in a good condition. Based on this, most of the models can locate the major parts of salient objects. According to Fig. 4, we can find that our model shown in Fig. 4 (a) achieves the best performance, which highlights the salient objects more complete and more accurate. For other models, we can find that the CNNs-based models (Fig. 4 (j)–(q)) achieve more promising results than the traditional non-deep learning models (Fig. 4 (e)–(i)), which also validates the effectiveness of deep learning techniques.

Subsequently, in 6*th* and 7*th* rows, the cases are more complex scenes, which exhibit low contrast between foreground and background, unclear depth as well as cluttered background. According to Fig. 4, it can be seen that our model performs best on the two images. In contrast, other models (Fig. 4 (e)–(q)) not only cannot pop-out the salient objects coarsely, but also highlight the background regions falsely. Among those models, especially some deep learning models such as [39,40] and [35], they yield poor results because of heavily depending on depth cues, which are in low quality. Lastly, for the

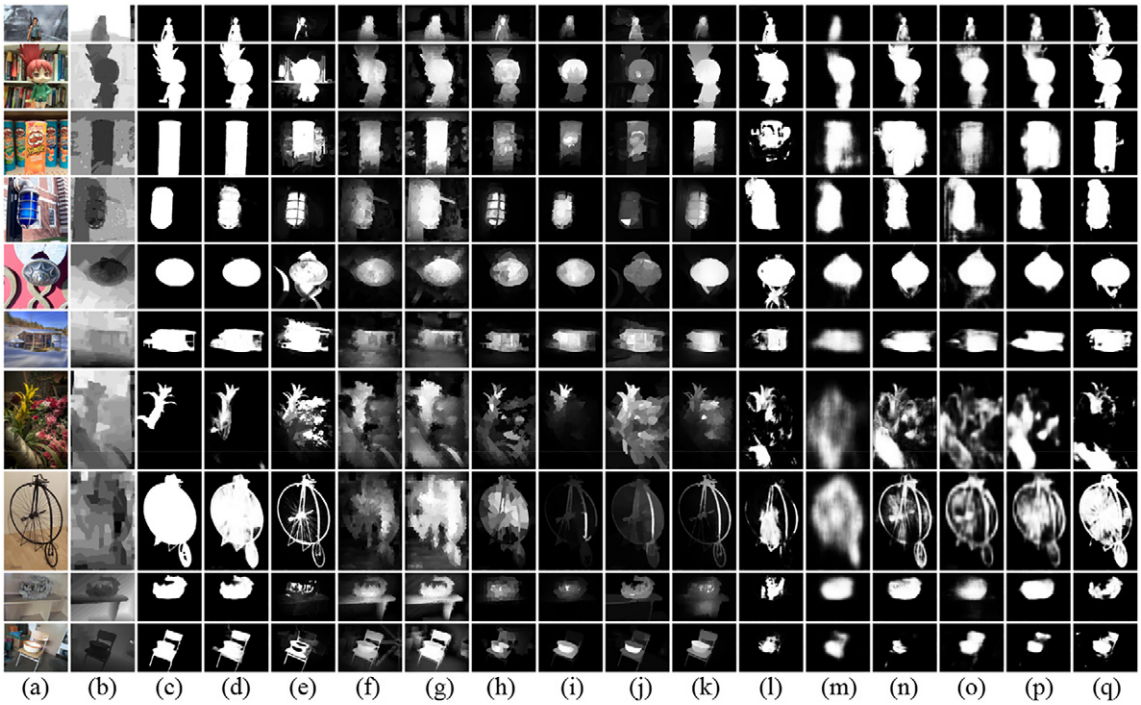

Fig. 4. Visualization comparison of different RGBD saliency models on several challenging scenes. (a): RGB, (b): Depth, (c): GT, (d): Ours, (e): ACSD, (f): LBE, (g): SE, (h): DCMC, (i): CDCP, (j): DF, (k): MDSF, (l): PCF, (m): CTMF, (n): AFNet, (o): CPFP, (p): MMCI, (q): TANet.

Table 2
Ablation studies are performed on three public RGBD datasets including NJU2K [22] and LFSD [55]. Noted, the best result in each column is marked in bold face.

		w/oAPP	w/oHHA	w/oAM	wHHADecoder	**Ours**
NJU2K-TE	$S\uparrow$	0.883	0.882	0.887	0.893	**0.893**
	$maxF\uparrow$	0.880	0.880	0.884	0.890	**0.891**
	$maxE\uparrow$	0.926	0.921	0.921	0.928	**0.930**
	$MAE\downarrow$	0.057	0.057	0.059	0.055	**0.055**
LFSD	$S\uparrow$	0.856	0.859	0.848	0.861	**0.876**
	$maxF\uparrow$	0.858	0.865	0.840	0.866	**0.877**
	$maxE\uparrow$	0.892	0.902	0.886	0.895	**0.912**
	$MAE\downarrow$	0.078	0.076	0.084	0.076	**0.070**

last two examples shown in 8*th*, 9*th* and 10*th* rows, they also present more complex seniors, such as the last example shows a washbasin on a chair, the 9*th* example shows a rare stone on a desk, and the 8*th* case presents a bicycle in a peculiar shape. Meanwhile, the three examples' depth maps are in low quality. Though they are challenging scenes, according to Fig. 4, we can find that our model (Fig. 4 (d)) still highlights the salient objects more completely and suppresses the background regions more effectively. The reason behind this can be owing to the sufficient and appropriate utilization of RGB and depth information via the bottom-up and top-down modules, where the attention maps are embedded to guide the estimation of salient objects and appearance information is given more concern.

4.3. Ablation studies

This section deeply analyses some vital components in our model via the quantitative and qualitative comparisons. Specifically, the vital components in our model contain the attention maps, the depth map HHA and the appearance information. Correspondingly, for studying the contribution of these components, we design several variations for our model, namely without the attention map, without the depth map HHA (our model only depends on RGB images) and without appearance information, which are denoted as "w/oAM", "w/oHHA" and "w/oAPP", respectively. Besides, similar to the usage of appearance information, we also explore the effect of the multi-level deep depth features in the top-down module (decoder part) and denote this variation as "wHHADecoder". To study the performance of these variations, *i.e.* " w/oAM", "w/oHHA", "w/oAPP" and "wHHADecoder", we also execute some comparisons from the perspective of quantitative and qualitative views, as shown in Table 2 and Fig. 5.

According to the quantitative comparison results shown in Table 2, we can find that our model outperforms other variations including w/oAM, w/oHHA, w/oAPP and wHHADecoder in terms of S-measure, max F-measure, max E-measure and MAE. Correspondingly, referring to the qualitative comparison results shown in Fig. 5, we can also see that our model shown in Fig. 5 (d) performs best, which not only completely pops-out the salient objects but also suppresses the background regions effectively. In contrast, most of the other variations shown in Fig. 5 (e)–(h) falsely highlight the background regions. For example, the top example is a yellow flower among the green plants and red flowers, which presents a cluttered background in the RGB image. It can be seen that the results of w/oAM, w/oHHA, w/oAPP shown in Fig. 5 (f)–(h) pop-out the red flowers mistakenly. In stark contrast, our model and wHHADecoder suppress the background regions successfully, as shown in Fig. 5 (d) and (e). However, the results of wHHADecoder falsely suppress the floral axis of yellow flower. For the bottom example (two people are talking), the results of wHHADecoder, w/oAM, w/oHHA and w/oAPP shown in Fig. 5 (e)–(h) falsely highlight the woman. In contrast, our model performs best as shown in Fig. 5 (d), which is the one nearest the ground truth shown in Fig. 5 (c).

The reasons behind this lie in that the depth and RGB information should not only be used sufficiently but also be utilized appropriately. As the aforementioned discussion, there are many challenging scenes for RGBD saliency detection. Meanwhile, if this scenes presented by RGB images further encounter low-quality depth maps, the RGBD saliency detection will become a tough challenging task. Thus, the sufficient and appropriate usage of depth and information is a crucial issue. Concretely, firstly, referring to these variations, w/oAM refers to our model without attention maps in both modules. Obviously, the resulting saliency maps shown in Fig. 5 (f) falsely highlight the background regions. As for w/oHHA, it means that the two-stream structure based bottom-up module in our model only contains RGB branch but still contains "AM", which is obtained based on depth map, and the top-down module stays the same. We can find that the results of w/oHHA shown in Fig. 5 (g) pop-out less background regions than w/oAM due to the retain of the attention maps. Therefore, through the two variations w/oAM and w/oHHA, we can validate the effectiveness of depth information in our model, and that is to say we should sufficiently and appropriately utilize the depth information.

Then, for wHHADecoder, it indicates that the top-down module in our model introduces the multi-level deep depth features. Following this settings, it can be seen that both the bottom-up and top-down modules all utilize the depth information. However, the discrimination of salient objects and background regions may be not clear in depth maps, such as the floral axis in the top example and the man in the bottom example shown in Fig. 5. Thus, we can say that the heavy trust on depth information may introduce disturbance cues for RGBD saliency detection. In contrast, for the two examples, the RGB images can differentiate the floral axis and the man via color contrast. On

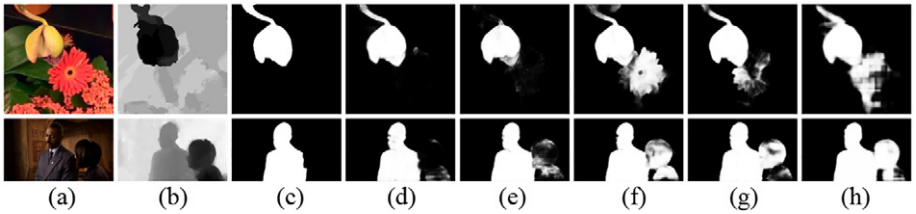

Fig. 5. Qualitative comparisons of several variations of the proposed RGBD saliency model on challenging scenes. (a): RGB, (b): Depth, (c): GT, (d): Ours, (e): wHHADecoder, (f): w/oAM, (g): w/oHHA, (h): w/oAPP.

the basis of this, our model not only uses the depth information adequately but also pays more attention to the RGB information, yielding high-quality saliency maps shown in Fig. 5 (d). Lastly, for w/oAPP, it is our model only with bottom-up module, which gives the equal status for RGB and depth information. Referring to Fig. 5 (h), we can find that its results detect the woman mistakenly. In sharp contrast, our model suppresses the woman effectively. Therefore, the two variations including wHHADecoder and w/oAPP not only validate that the employment of depth information should also be appropriate, but also demonstrate the effectiveness of appearance information.

Generally speaking, according to the aforementioned ablation studies, we can find that the usage of RGB and depth information is very important for RGBD saliency detection, and both cues should be used in a sufficient and appropriate way. Meanwhile, all these quantitative and qualitative results also demonstrate the rationality and effectiveness of the design of our model.

5. Conclusion

This paper proposes a novel RGBD saliency model, which contains bottom-up and top-down modules, to perform saliency detection on RGBD scenes. The core issue of our model lies in that the utilization of the RGB and depth information should be in a sufficient and appropriate way. During the entire process, we deploy attention maps to boot the salient objects' location and give more concern to the appearance information, which further strengthen the combination of both modules. Specifically, we first introduce the attention maps to both modules for guiding the differentiation of salient objects and background regions, which sufficiently utilize the depth information and effectively ease the deficiency of low-quality depth maps. Secondly, we further pay more attention to appearance information, *i.e.* the multi-level deep RGB features, via the top-down module, which adequately employs the appearance information by gradually aggregating with the fused deep feature generated by bottom-up module in a coarse-fine way. In this way, we can obtain a high-quality saliency map for each RGBD image. Extensive experiments are performed on five public and challenging RGBD datasets, and the results show that our model consistently outperforms the state-of-the-art RGBD saliency models, which further demonstrates the effectiveness and superiority of our model.

CRediT authorship contribution statement

Xiaofei Zhou: Conceptualization, Methodology, Writing - original draft. **Gongyang Li:** Methodology, Software, Validation. **Chen Gong:** Resources, Writing - review & editing. **Zhi Liu:** Writing - review & editing, Supervision, Funding acquisition. **Jiyong Zhang:** Funding acquisition, Project administration.

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 was supported by the National Natural Science Foundation of China under grants 61901145, 61771301, 61973162 and 61972123.

References

[1] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E.K. Fishman, A.L. Yuille, Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation, Computer Vision and Pattern Recognition, CVPR, IEEE. 2018, pp. 8280–8289.
[2] P. Mukherjee, B. Lall, Saliency and KAZE features assisted object segmentation, Image Vis. Comput. 61 (2017) 82–97.
[3] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, W. Zhang, Saliency-guided quality assessment of screen content images, IEEE Trans. Multimedia 18 (6) (2016) 1098–1110.
[4] D. Stefic, I. Patras, Action recognition using saliency learned from recorded human gaze, Image Vis. Comput. 52 (2016) 195–205.
[5] R. Quispe, H. Pedrini, Improved person re-identification based on saliency and semantic parsing with deep neural network models, Image Vis. Comput. 92 (2019) 103809.
[6] C. Gong, D. Tao, W. Liu, S.J. Maybank, M. Fang, K. Fu, J. Yang, Saliency propagation from simple to difficult, Computer Vision and Pattern Recognition, CVPR, IEEE. 2015, pp. 2531–2539.
[7] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, S. Li, Salient object detection: a discriminative regional feature integration approach, Computer Vision and Pattern Recognition, CVPR, IEEE. 2013, pp. 2083–2090.
[8] N. Tong, H. Lu, X. Ruan, M.-H. Yang, Salient object detection via bootstrap learning, Computer Vision and Pattern Recognition, CVPR, IEEE. 2015, pp. 1884–1892.
[9] X. Zhou, Z. Liu, G. Sun, L. Ye, X. Wang, Improving saliency detection via multiple kernel boosting and adaptive fusion, IEEE Signal Process Lett. 23 (4) (2016) 517–521.
[10] X. Zhou, Z. Liu, C. Gong, L. Wei, Improving video saliency detection via localized estimation and spatiotemporal refinement, IEEE Trans. Multimedia 20 (11) (2018) 2993–3007.
[11] P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, Learning uncertain convolutional features for accurate saliency detection, International Conference on Computer Vision, ICCV, IEEE. 2017, pp. 212–221.
[12] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, A. Borji, Detect globally, refine locally: a novel approach to saliency detection, Computer Vision and Pattern Recognition, CVPR, IEEE. 2018, pp. 3127–3135.
[13] G. Li, Y. Xie, T. Wei, K. Wang, L. Lin, Flow guided recurrent neural encoder for video salient object detection, Computer Vision and Pattern Recognition, CVPR, IEEE. 2018, pp. 3243–3252.
[14] H. Wen, X. Zhou, Y. Sun, J. Zhang, C. Yan, Deep fusion based video saliency detection, J. Vis. Commun. Image Represent. 62 (2019) 279–285.
[15] W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, Computer Vision and Pattern Recognition, CVPR, IEEE. 2019, pp. 1448–1457.
[16] Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, Computer Vision and Pattern Recognition, CVPR, IEEE. 2018, pp. 6154–6162.
[17] C. Sun, D. Wang, H. Lu, M.-H. Yang, Learning spatial-aware regressions for visual tracking, Computer Vision and Pattern Recognition, CVPR, IEEE. 2018, pp. 8962–8970.
[18] C.P. Lau, C.P. Yung, L.M. Lui, Image retargeting via Beltrami representation, IEEE Trans. Image Process. 27 (12) (2018) 5787–5801.
[19] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, Computer Vision and Pattern Recognition, CVPR, IEEE. 2012, pp. 454–461.
[20] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, Proceedings of International Conference on Internet Multimedia Computing and Service, ICIMCS, ACM. 2014, pp. 23–27.
[21] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: a benchmark and algorithms, European Conference on Computer Vision, ECCV, Springer. 2014, pp. 92–109.
[22] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center–surround difference, International Conference on Image Processing, ICIP, IEEE. 2014, pp. 1115–1119.
[23] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, W. Lin, Saliency detection for stereoscopic images, IEEE Trans. Image Process. 23 (6) (2014) 2625–2636.
[24] R. Ju, Y. Liu, T. Ren, L. Ge, G. Wu, Depth-aware salient object detection using anisotropic center-surround difference, Signal Process. Image Commun. 38 (2015) 115–126.
[25] D. Feng, N. Barnes, S. You, C. McCarthy, Local background enclosure for RGB-D salient object detection, Computer Vision and Pattern Recognition, CVPR, IEEE. 2016, pp. 2343–2350.
[26] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, C. Hou, Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion, IEEE Signal Process Lett. 23 (6) (2016) 819–823.
[27] J. Guo, T. Ren, J. Bei, Salient object detection for RGB-D image via saliency evolution, International Conference on Multimedia and Expo, ICME, IEEE. 2016, pp. 1–6.
[28] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, T. Ren, Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning, IEEE Trans. Image Process. 26 (9) (2017) 4204–4216.
[29] A. Wang, M. Wang, RGB-D salient object detection via minimum barrier distance transform and saliency fusion, IEEE Signal Process Lett. 24 (5) (2017) 663–667.
[30] C. Zhu, G. Li, W. Wang, R. Wang, An innovative salient object detection using center-dark channel prior, International Conference on Computer Vision, ICCV, IEEE. 2017, pp. 1509–1515.
[31] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, Q. Yang, RGBD salient object detection via deep fusion, IEEE Trans. Image Process. 26 (5) (2017) 2274–2285.
[32] R. Shigematsu, D. Feng, S. You, N. Barnes, Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features, International Conference on Computer Vision, ICCV, IEEE. 2017, pp. 2749–2757.

[33] X. Xu, Y. Li, G. Wu, J. Luo, Multi-modal deep feature learning for RGB-D object detection, Pattern Recogn. 72 (2017) 300–313.
[34] J. Han, H. Chen, N. Liu, C. Yan, X. Li, CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion, IEEE Trans. Cybern. 48 (11) (2017) 3171–3183.
[35] H. Chen, Y. Li, Progressively complementarity-aware fusion network for RGB-D salient object detection, Computer Vision and Pattern Recognition, CVPR, IEEE. 2018, pp. 3051–3060.
[36] Z. Liu, S. Shi, Q. Duan, W. Zhang, P. Zhao, Salient object detection for RGB-D image by single stream recurrent convolution neural network, Neurocomputing 363 (2019) 46–57.
[37] H. Chen, Y. Li, D. Su, Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection, IEEE Trans. Cybern. (2019) https://doi.org/10.1109/TCYB.2019.2934986.
[38] C. Zhu, X. Cai, K. Huang, T.H. Li, G. Li, Pdnet: prior-model guided depth-enhanced network for salient object detection, International Conference on Multimedia and Expo, ICME, IEEE. 2019, pp. 199–204.
[39] N. Wang, X. Gong, Adaptive fusion for RGB-D salient object detection, IEEE Access 7 (2019) 55277–55284.
[40] H. Chen, Y. Li, D. Su, 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.
[41] H. Chen, Y. Li, Three-stream attention-aware network for RGB-D salient object detection, IEEE Trans. Image Process. 28 (6) (2019) 2825–2835.
[42] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, L. Zhang, Contrast prior and fluid pyramid integration for RGBD salient object detection, Computer Vision and Pattern Recognition, CVPR, IEEE. 2019, pp. 3927–3936.
[43] A. Borji, Boosting bottom-up and top-down visual features for saliency estimation, Computer Vision and Pattern Recognition, CVPR, IEEE. 2012, pp. 438–445.
[44] H. Tian, Y. Fang, Y. Zhao, W. Lin, R. Ni, Z. Zhu, Salient region detection by fusing bottom-up and top-down features extracted from a single image, IEEE Trans. Image Process. 23 (10) (2014) 4389–4398.
[45] N. Liu, J. Han, M.-H. Yang, PiCANet: learning pixel-wise contextual attention for saliency detection, Computer Vision and Pattern Recognition, CVPR, IEEE. 2018, pp. 3089–3098.
[46] J. Fu, H. Zheng, T. Mei, Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition, Computer Vision and Pattern Recognition, CVPR, IEEE. 2017, pp. 4438–4446.
[47] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representations, 2014. pp. 1–14.
[48] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-D images for object detection and segmentation, European Conference on Computer Vision, ECCV, Springer. 2014, pp. 345–360.
[49] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[51] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: aggregating multi-level convolutional features for salient object detection, International Conference on Computer Vision, ICCV, IEEE. 2017, pp. 202–211.
[52] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P.H. Torr, Deeply supervised salient object detection with short connections, Computer Vision and Pattern Recognition, CVPR, IEEE. 2017, pp. 3203–3212.
[53] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, Proceedings of the 22*nd* ACM International Conference on Multimedia, MM, ACM. 2014, pp. 675–678.
[54] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the 13*th* International Conference on Artificial Intelligence and Statistics, 9, 2010. pp. 249–256.
[55] N. Li, J. Ye, Y. Ji, H. Ling, J. Yu, Saliency detection on light field, Computer Vision and Pattern Recognition, CVPR, IEEE. 2014, pp. 2806–2813.
[56] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: a new way to evaluate foreground maps, International Conference on Computer Vision, ICCV, IEEE. 2017, pp. 4548–4557.
[57] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, Proceedings of the 27*th* International Joint Conference on Artificial Intelligence, IJCAI. 2018, pp. 698–704.
[58] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks, IEEE Trans. Image Process. 27 (1) (2018) 38–49.