Image and Vision Computing 101 (2020) 103964

Contents lists available at ScienceDirect

Image and Vision Computing

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

Cross-modal feature extraction and integration based RGBD saliency detection

Liang Pan [a], Xiaofei Zhou [a,*], Ran Shi [b], Jiyong Zhang [a], Chenggang Yan [a]

[a] School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China
[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

ARTICLE INFO

Article history:
Received 25 May 2020
Accepted 31 May 2020
Available online 28 June 2020

Keywords:
RGBD
Saliency
Cross-modal
Feature extraction
Integration

ABSTRACT

In RGBD saliency detection research field, RGB and depth cues are generally given the same status by RGBD saliency models. However, they ignore that both modalities are significantly different in inherent attribution so that effective features cannot be drawn from depth maps. In order to address this issue, this paper proposes a novel RGBD saliency model including two key components: the contrast-guided depth feature extraction (CDFE) module and the cross-modal feature integration (CFI) module. Specifically, considering the specific properties of depth information, we first design a targeted CDFE module, which learns multi-level deep depth features by strengthening the depth contrast between foreground and background, to provide multi-level deep depth features. Then, to sufficiently and reasonably integrate multi-level cross-modal features, namely the multi-level deep RGB and depth features, we equip the saliency inference branch with the CFI module, which contains two successive steps, *i.e.* information enrichment and feature enhancement. Extensive experiments are conducted on five challenging RGBD datasets, and the experimental results clearly demonstrate the effectiveness and superiority of the proposed model against the state-of-the-art RGBD saliency models.

1. Introduction

Salient object detection, which aims to identify the most visually attractive objects in a scene, is one of the fundamental tasks in computer vision area. It can be broadly deployed to many visual tasks such as segmentation [1], video compression [2], person re-identification [3] and human action recognition [4]. Recently, lots of RGB saliency models including traditional methods [5–8] and convolutional neural networks (CNNs) based methods [9–12] have been proposed to boost the performance of salient object detection. However, the existing RGB saliency models may generate unsatisfactory saliency maps when dealing with challenging situations, such as similar appearance between salient objects and background, cluttered background, heterogeneous salient objects and so on. Fortunately, with the rapid development of stereo sensors such as Microsoft Kinect, RealSense and the corresponding depth estimation algorithms [13,14] can conveniently provide important supplementary information such as 2D locations, structures and 3D layouts of objects. Thus, on basis of this, we can significantly improve the performance of saliency models by jointly employing RGB and depth information to perform saliency detection.

To exploit depth cues, the first issue is how to effectively extract features from depth maps. The prior traditional RGBD models [15–18] mainly use domain knowledge to design hand-crafted features, whose representation ability is limited due to the lack of high-level information. Meanwhile, with the demonstration of CNNs' superiority, lots of CNNs based RGBD models [19–26] are presented to automatically learn deep depth features from the depth maps by employing the widely used feature extraction networks [27,28]. However, most of these RGBD saliency models extract their deep depth features by deploying the pre-trained models VGG-16 [27] or ResNet [28], which are learned from a large scale RGB image dataset ImageNet, and some of these RGBD saliency models even employ a Siamese architecture, sharing the parameters in RGB and depth branches. Obviously, the inherent attribution of depth and RGB information is considerably different, where depth maps don't contain the same abundant color and texture information as RGB images. Therefore, no matter performing fine-tuning on the pre-trained models or adopting the Siamese architecture may lead to the ambiguity in depth feature extraction. Successively, the second issue of RGBD saliency detection, *i.e.* how to effectively integrate RGB and depth features, should also be given more concern. Certainly, there are also many efforts have been devoted to explore the fusion mechanism of both modalities. For example, the early efforts [15–18] try to employ linear summation to fuse the RGB and depth information. For the CNNs based RGBD saliency models [20,21,23], they design symmetry architecture to make interactions between RGB features and depth features. Through an in-depth analysis, it can be found that the fusion strategy treats the deep RGB and depth features

[*] Corresponding author.
E-mail address: zxforchid@outlook.com (X. Zhou).

https://doi.org/10.1016/j.imavis.2020.103964
0262-8856/

equally. However, as the aforementioned, the inherent attribution of the two modalities is considerably different. Thus, in recent works [24–26], they try to fuse the RGB features and depth features by distinguish the two modalities, namely paying more attention to depth information or RGB information. Following this way, the performance of RGBD saliency detection are push forward largely. But when dealing with some challenging scenes, the performance of existing cutting-edge models will also decline to some degree.

Motivated by the aforementioned analysis, we propose a novel RGBD saliency model, as shown in Fig. 1, to perform salient object detection in RGBD scenes. Firstly, taking a full account of the differences between RGB and depth modalities, we deploy a more targeted strategy, *i.e.* the contrast-guided depth feature extraction (CDFE) module, as the depth branch shown in Fig. 1, which is significantly different from the RGB branch. In the CDFE module, the obtained multi-level deep depth features can not only supply sufficient low-level spatial details but also imply high-level depth contexts. Meanwhile, we also introduce the contrast loss [25] to increase the depth contrast between foreground and background, that is to say the obtained deep depth features are given the ability to discriminate the salient objects and background to some extent. Here, we should note that the key difference between our model and CPFP [25] is that CPFP employs the enhanced depth map, *i.e.* the output of the depth branch, as the depth feature, while our model directly adopts the outputs of some convolutional layers in depth branch as the multi-level deep depth features.

Then, to sufficiently integrate the multi-level deep RGB and depth features, we explore the cross-modal complementarity and design a cross-modal feature integration (CFI) module, which consists of two successive steps, *i.e.* information enrichment (IE) and feature enhancement (FE). On basis of this, we leverage six CFI modules to construct the saliency inference branch. Concretely, we first exploit the IE step to provide additional depth features, which are progressively combined with multi-level deep RGB features. Then, to further improve the representation ability of deep features, we perform the FE step, which takes the advantage of cross-modal features, to pop-out salient regions and suppress background noises. Following this way, we can integrate cross-modal deep features sufficiently and appropriately, achieving high-performance salient object detection in RGBD scenes.

Overall, the main contributions can be concluded as follows:

1. Considering the inherent difference between the depth and RGB information, we propose a novel RGBD saliency model, which contains two key components including contrast-guided depth feature extraction (CDFE) module and cross-modal feature integration (CFI) module.
2. To sufficiently and reasonably utilize the two modalities, the CDFE module can provide multi-level deep depth features, involving low-level spatial details and high-level depth contexts. Then, the CFI module is deployed to integrate the two modalities by information enrichment and feature enhancement.
3. Extensive experiments conducted on five public RGBD datasets show the effectiveness and superiority of the proposed RGBD saliency model against fifteen state-of-the-art RGBD saliency models.

2. Related works

The early efforts [15–18,29–31] on RGBD saliency detection try to identify salient objects by designing various hand-crafted features, such as multi-contextual contrast [15], anisotropic center-surround difference [16], global priors [17], local background enclosure [18], center-dark channel prior [30] and so on. Specifically, in [15], Peng et al. leverage low-level feature contrast, mid-level region grouping and high-level priors enhancement to perform multi-stage saliency prediction. In [17], global priors, which are divided into depth, background and orientation priors, are combined with region contrast to produce saliency maps. In [18], to solve the challenge scene, *i.e.* salient regions are with low depth contrast, Feng et al. combine angular density and gap to construct the local background enclosure feature. In general, we can find that the hand-crafted features based RGBD saliency models are deficient in representing the salient objects due to the lack of high-level context features.

Recently, with the rapid development of convolutional neural networks (CNNs), many RGBD saliency models [19–26,32–34] also try to employ the deep learning techniques, yielding high-quality saliency maps. For example, in [32], a simple CNNs model is first employed to fuse hand-crafted RGBD features, and then a Laplacian Propagation is adopted as a post-processing step to further promote the performance. In [33], Shigematsu et al. use two independent CNNs models to process RGB images and hand-crafted depth features. Although the two

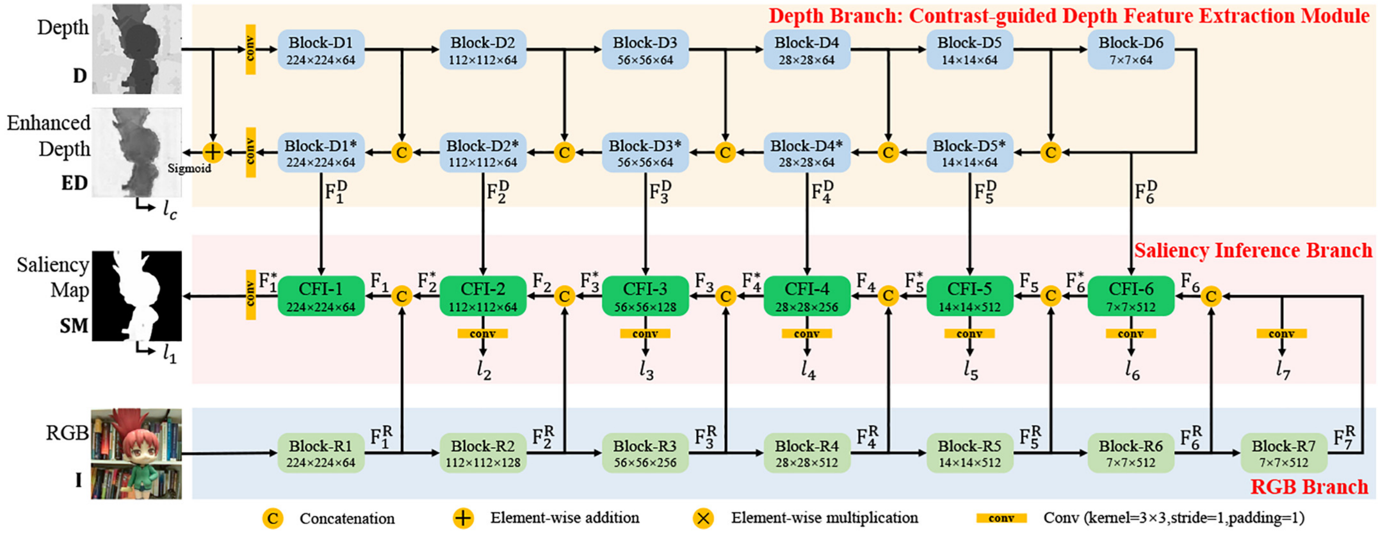


Fig. 1. Illustration of the proposed RGBD saliency model. Firstly, the depth branch, namely the contrast-guided depth feature extraction (CDFE) module, which contains Block-D*i* ($i = 1, \ldots, 6$) and Block-D*i** ($i = 1, \ldots, 5$), is employed to extract multi-level depth features $\{\mathbf{F}_i^D\}_{i=1}^6$. Meanwhile, the RGB branch, which includes Block-R*i* ($i = 1, \ldots, 7$), is deployed to extract multi-level RGB features $\{\mathbf{F}_i^D\}_{i=1}^7$. Then, the saliency inference branch, which is equipped with cross-modal feature integration modules CFI-*i* ($i = 1, \ldots, 6$), is used to progressively integrate multi-level cross-modal features and produce a high-quality saliency map **SM**. Notably, before concatenation operation, we exploit a bilinear upsampling layer for the sake of feature maps' resolution.

pioneering RGBD saliency models achieve better performance than the early traditional RGBD saliency models, the high-level deep features is extracted from well-designed initial depth features, which is still can't provide satisfactory performance.

To automatically obtain both low-level and high-level features, outstanding feature extraction networks [27,28] are widely used as backbone to build a two-stream architecture including RGB and depth branches. For example, in [19], Han et al. first leverage cross-view transfer and cross-view late fusion to initialize the depth branch, and then the two branches jointly predict saliency maps, respectively. In [22], Chen et al. incorporate the global understanding and local capturing modules into the VGG-based RGB and depth branches, which are combined by multiple paths and cross-modal interactions. In [20], a complementarity-aware fusion module is designed to sufficiently dig the cross-modal features. In [21], Chen et al. exploit channel-wise attention mechanism to determine the contribution of cross-modal features in a top-down inference manner. In [23], Chen et al. deploy an additional cross-modal distillation stream to form a three-stream architecture, which explores the complementarity in both bottom-up feature extraction and top-down saliency inference. In [24], residual connections are used to extract and fuse multi-level cross-modal features and a depth-induced multi-scale weighting module is then proposed to improve the discrimination. Overall, the aforementioned CNNs based RGBD saliency models mainly try to explore the cross-modal information complementarity.

Besides, there are also some works [25,26,34] pay attention to the negative influence of low-quality depth maps. For example, in [25], Zhao et al. propose a contrast-enhanced network to enhance the contrast in depth maps. In [34], Fan et al. propose a depth depurator to discard low-quality depth maps. In [26], Zhou et al. give more concern on the appearance information in RGB features during the saliency inference process. In conclusion, most CNNs based saliency models directly use ImageNet pre-trained model as backbone, which is used to extract features from depth maps. However, they often ignore the inherent attribution difference between two modalities. In this paper, we propose a contrast-guided depth feature extraction module, which can supply multi-level depth features with sufficient spatial details and depth contexts. Meanwhile, we also design a cross-modal feature integration module to combine RGB and depth features effectively.

3. The proposed method

In this section, we first briefly introduce the overall architecture of the proposed RGBD saliency model in Section 3.1. Then, we describe the contrast-guided depth feature extraction (CDFE) module and cross-modal feature integration (CFI) module in detail in Sections 3.2 and 3.3, respectively. Lastly, the model training and implementation details are given in Section 3.4.

3.1. Overall architecture

As shown in Fig. 1, the proposed model mainly contains three components including the depth branch, namely contrast-guided depth feature extraction (CDFE) module, the RGB branch as well as the saliency inference branch which is equipped with the cross-modal feature integration (CFI) module. Concretely, the input of our model is RGB image **I** and depth map **D**. Firstly, referring to [12], we modify ResNet [28] to form the RGB branch, which contains seven blocks Block-Ri ($i = 1, \ldots, 7$), yielding the multi-level deep RGB features $\{\mathbf{F}_i^R\}_{i=1}^7$. Secondly, for the depth branch, we design the contrast-guided depth feature extraction (CDFE) module to extract the multi-level deep depth features $\{\mathbf{F}_i^D\}_{i=1}^6$. Certainly, we can also gain the enhanced depth map **ED**. Lastly, the saliency inference branch is designed in a decoder way, where we configure a set of cross-modal feature integration (CFI) modules, *i.e.* CFI-i ($i = 1, \ldots, 6$), to progressively integrate the multi-level deep RGB and depth features. Following this way, we can obtain the high-quality RGBD saliency map **SM**.

3.2. Contrast-guided depth feature extraction module (CDFE)

Generally, depth maps are capable of discriminating the foreground and background to some extent. Meanwhile, due to the existence low-

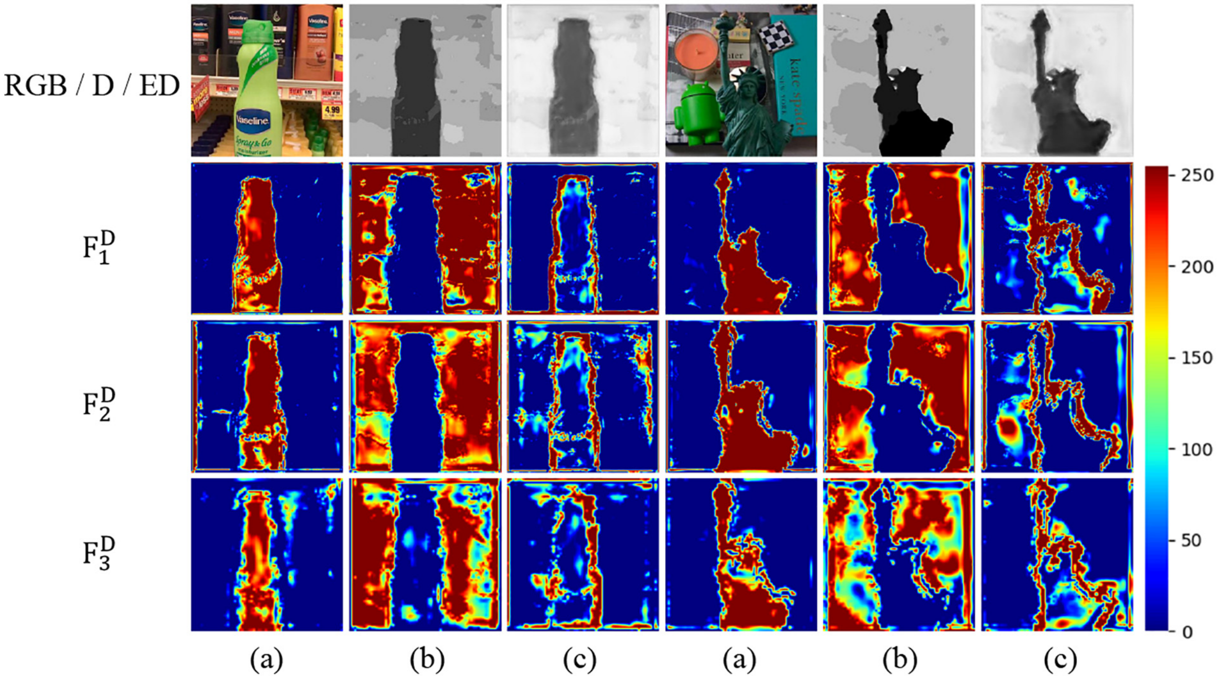


Fig. 2. Visualization results of some feature maps in low-level depth features *i.e.* $\{\mathbf{F}_i^D\}_{i=1}^3$: for each example, the first row has a RGB image, a depth map and an enhanced depth map from left to right. The rest of rows represent three categories of feature maps in $\mathbf{F}_1^D$, $\mathbf{F}_2^D$, $\mathbf{F}_3^D$, respectively. Notably, we scale all feature maps to the same resolution. Comparing with the enhanced depth map **ED**, (a) and (b) have higher discrimination between foreground and background, and (c) indicates the contrast on boundary.

quality depth map, the integration of deep RGB and depth features may degenerate the performance of RGBD saliency detection. To solve this problem, in [25], the CPFP model designs the contrast-enhanced network (CEN) to improve the depth map quality. Then, the CPFP directly uses the enhanced depth map to elevate the RGB features' discrimination ability, *i.e.* distinguishing salient objects and background. However, the method ignores the abundant rich deep depth features within the layers of CEN, where the deep depth features not only provide spatial details but also contain rich context information. Besides, inspired by the successful deployment of residual structure in saliency maps' refinement in [12], the depth branch also adopts a similar architecture and is designed as the contrast-guided depth feature extraction (CDFE) module exhibited in Fig. 1.

Specifically, the CDFE module adopts the encoder-decoder architecture, which contains six encoder blocks Block-Di ($i = 1, \dots, 6$) and five decoder blocks Block-Di^* ($i = 1, \dots, 5$). Among these blocks, each block has one convolutional layer (kernel size $= 3 \times 3$, stride $= 1$, padding $= 1$) followed by a batch normalization layer and a ReLU layer. Besides, for the last five encoder blocks Block-Di ($i = 2, \dots, 6$), each of them also employs an additional max pooling layer (kernel size $= 2 \times 2$, stride $= 2$, padding $= 0$). Meanwhile, to increase the depth map **D**'s channel number (64) and generate the one-channel output, we employ one convolutional layer (kernel size $= 3 \times 3$, stride $= 1$, padding $= 1$) at the beginning and end of the CDFE module, respectively. Further, we exploit a sigmoid function to scale the additive result to [0, 1] and obtain the final enhanced depth map **ED**. Following this way, we can obtain abundant deep depth features $\{\mathbf{F}_i^D\}_{i=1}^6$. According to this description and the detailed structures shown in Fig. 1, we can easily differentiate the difference between the CEN and our CDFE.

In particular, we give a visualization comparison in Fig. 2. Concretely, we can observe that the feature maps in low-level depth features $\{\mathbf{F}_i^D\}_{i=1}^3$ preserve complete spatial structures and are with high contrast between foreground and background, which is more distinct than the depth map **D** and even the enhanced depth map **ED**. In addition, some low-level features also preserve the contrast on boundaries. Further, for the high-level depth features $\{\mathbf{F}_i^D\}_{i=4}^6$, their feature maps' spatial structures gradually degrade, and they mainly store high-level context information. In conclusion, the proposed CDFE module can provide multi-level deep depth features by increasing the depth contrast between foreground and background, where both low-level spatial contrasts and high-level depth contexts are captured.

3.3. Cross-modal feature integration module (CFI)

Through deploying the contrast-guided depth feature extraction (CDFE) module, we can obtain multi-level deep depth features $\{\mathbf{F}_i^D\}_{i=1}^6$. Thus, to effectively utilize the depth features, we design the cross-modal feature integration (CFI) module shown in Fig. 3 and deploy six CFI-i ($i = 1, \dots, 6$) to the saliency inference branch, where the CFI module contains two successive steps including information enrichment (IE) and feature enhancement (FE).

Concretely, most saliency models commonly suffer the problem that is lack of discrimination in some challenging scenes, such as low-contrast between foreground and background, complex background, multiple objects and so on. Fortunately, the abundant spatial contrast information can be captured by low-level depth features shown in Fig. 2, thus we design the IE step to effectively improve the discrimination ability, where we employ a concatenation operation to simply combine each level deep depth feature with the corresponding deep RGB feature. Following this way, we can not only leverage diverse spatial contrasts on foreground, background and boundary in low-level depth features, but also make full use of depth contexts in high-level depth features, yielding more discriminative RGBD saliency model. Thus, we can define the overall IE step as follows:

$$\mathbf{F}_i^C = \left[Conv1(\mathbf{F}_i), \mathbf{F}_i^D\right], \tag{1}$$

where $\mathbf{F}_i^D$ denotes the i-th level depth feature, $\mathbf{F}_i^D$ together with $\mathbf{F}_i$ are the input of the CFI-i module, and $\mathbf{F}_i^C$ denotes the output of the IE step. Generally, we first use a convolutional layer Conv-1, denoted by *Conv1*, to reduce the channel number of $\mathbf{F}_i$. Then, we obtain the $\mathbf{F}_i^C$ by using a concatenation operation represented as [,]. Notably, our information enrichment (IE) step refers to the concatenation operation and doesn't contain the convolution layer Conv-1, as shown in Fig. 3.

Then, we employ two convolutional layers {Conv-2, Conv-3} to further process $\mathbf{F}_i^C$, yielding the deep feature $\mathbf{F}_1{}'$. Successively, to further improve the representation of feature maps in $\mathbf{F}_i{}'$, we perform the feature enhancement (FE) step. Specifically, we incorporate the depth feature $\mathbf{F}_i^D$ again, which is used to jointly integrate with the deep feature $\mathbf{F}_i{}'$ *via* convolutional layers Conv-4 and {Conv-5, Conv-6} and a sigmoid layer, yielding the refinement feature $\mathbf{RF}_i$. Through this way, we can take full advantage of deep depth features, which not only take part in the generation of fused deep feature but also make a refinement for the integration of the cross-modalities cues. Lastly, we regard the refinement feature $\mathbf{RF}_i$ as an attention map and make an integration with the deep feature $\mathbf{F}_i{}'$, yielding the final enhanced deep feature $\mathbf{F}_i{}^*$. The overall process can be defined as:

$$\begin{aligned}
&\mathbf{RF}_i = sigmoid\left(Conv6\left(Conv5\left(\left[Conv4(\mathbf{F}_i'), \mathbf{F}_i^D\right]\right)\right)\right) \\
&\mathbf{F}_i^* = \mathbf{F}_i'\odot\mathbf{RF}_i + \mathbf{F}' \\
&\mathbf{SM} = Conv(\mathbf{F}_1^*)
\end{aligned}, \tag{2}$$

where **SM** means the final saliency map, $\mathbf{F}_i^D$ denotes the i-th level deep depth feature, $\mathbf{F}_i^*$ denotes enhanced deep feature, which is also the final output of the CFI-i module, and $\mathbf{RF}_i$ denotes the refinement feature. Here, the convolutional layer "conv" after the CFI-1 module show in Fig. 3 is denoted as Conv, and the convolutional layer Conv-4 is used to reduce the channel number of $\mathbf{F}_i{}'$ to 64, and ⊙ means element-wise multiplication. Besides, to preliminary validate the effect of FE step, we lock on two channels $\{\mathbf{FM}_1, \mathbf{FM}_2\}$ in deep feature $\mathbf{F}_1{}'$ and observe

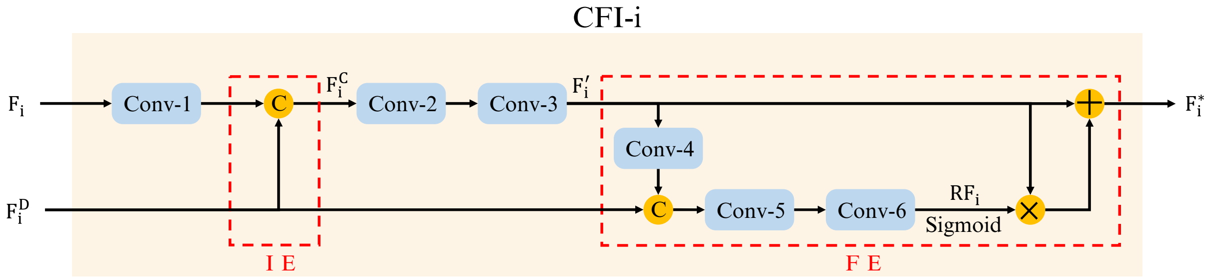


Fig. 3. Illustration of the cross-modal feature integration (CFI) module.

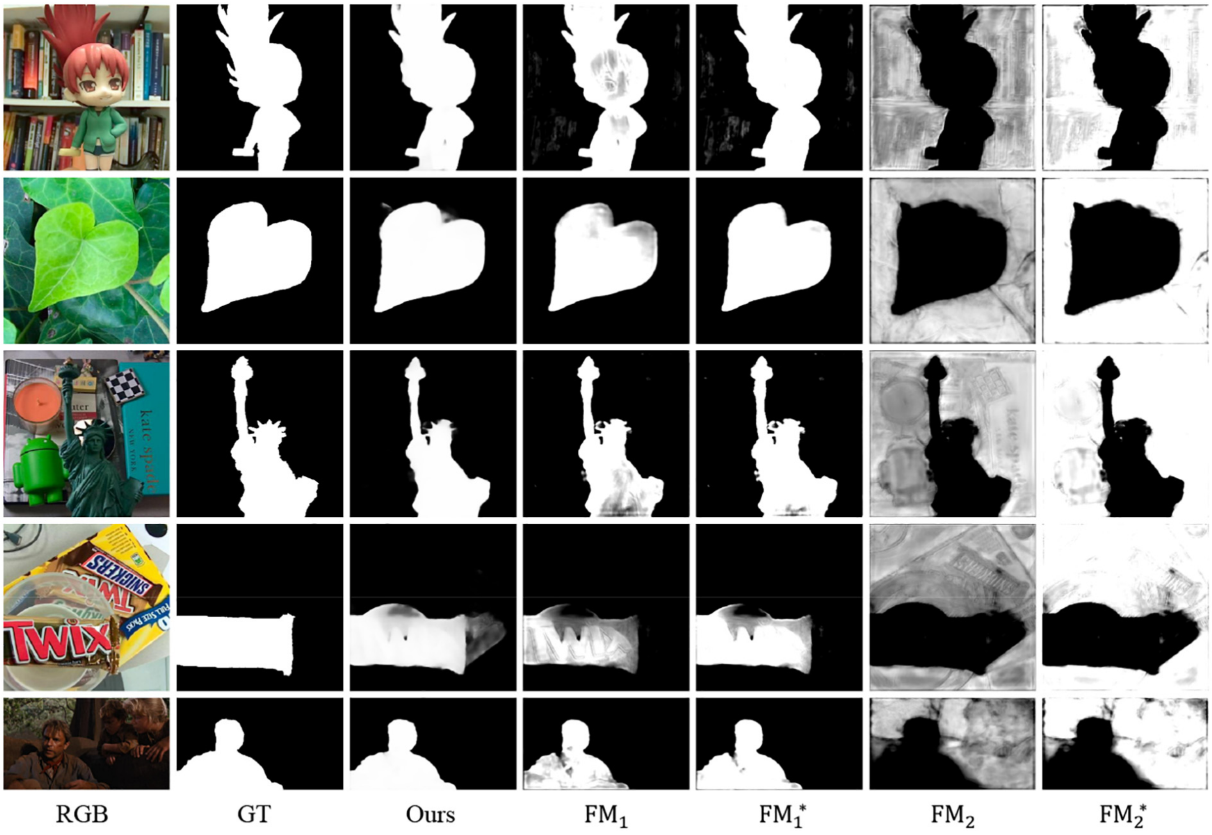


Fig. 4. Visualization comparisons over feature maps before and after the FE step. $\mathbf{FM}_1$ and $\mathbf{FM}_2$ are feature maps in feature $\mathbf{F}_1'$, which has not been processed. $\mathbf{FM}_1^*$**and** $\mathbf{FM}_2^*$ are corresponding feature maps in feature $\mathbf{F}_1^*$, which has been processed by the FE step.

their changes after the FE step, as shown in Fig. 4. We can find that the FE step can not only refine salient regions ($\mathbf{FM}_1$ and $\mathbf{FM}_1^*$), but also filter the noises of background ($\mathbf{FM}_2$ and $\mathbf{FM}_2^*$).

3.4. Model learning and implementation

3.4.1. Loss functions

To train the contrast-guided depth feature extraction (CDFE) module, we also employ a contrast loss [25]. Besides, we conduct deep supervision strategy [35] on CFI-*i* ($i = 1, \ldots, 6$) and Block-R7. Thus, the total loss $\mathscr{L}$ can be denoted as:

$$L = \sum_{i=1}^{7} l_i + l_c, \tag{3}$$

where l_i refers to the hybrid loss [12] including BCE, SSIM and IoU losses, and l_c refers to the contrast loss.

3.4.2. Implementation setups

We implement our RGBD saliency model using Pytorch 1.2.0 framework [36] on a NVIDIA GeForce RTX 2080Ti GPU (with 11 GB memory). We first initialize blocks Block-R*i* ($i = 1, 2, 3, 4$) by using the ImageNet pre-trained ResNet-34 [28], and the parameters in rest blocks are initialized by Xavier [37]. Then, the Adam optimizer [38] is used to minimize the total loss shown in Eq. (3), where we set betas, eps, weigh decay as $(0.9, 0.999)$, 10^{-8}, 0, respectively. Besides, we train our model for totally 18,000 iterations and the batch size is set to 8. Notably, the learning rate of the contrast-guided depth feature extraction (CDFE) module is set to 10^{-6}, and the learning rates of the RGB branch and the saliency inference branch are initialized with 10^{-4}, which are divided by 10 after 13,000 iterations.

3.4.3. Training dataset

We choose the same training data as [25], in which the training data consist of 1400 image pairs from NJU2K [16] and 650 image pairs from NLPR [15]. Besides, we also augment the training dataset by horizontal flipping and rotation with angles 90°, 180°, and 270° to 10,250 image pairs. Meanwhile, for training the proposed RGBD saliency model, each input RGBD image pair is resized to 256 × 256.

4. Experimental results

In this section, we will first give a brief introduction for the RGBD datasets and evaluation metrics in Section 4.1. Then, in Section 4.2, we will quantitatively and qualitatively make a comprehensive comparison for our model and the state-of-the-art RGBD saliency models. Lastly, the detailed ablation analysis will be presented in Section 4.3.

4.1. Datasets and evaluation metrics

4.1.1. Datasets

We evaluate the performance of the proposed model by performing extensive experiments on five RGBD datasets including NJU2K [16], NLPR [15], STEREO [39], LFSD [40] and DES [41]. In the following, we will give a brief introduction of these datasets. **NJU2K** contains 2003 image pairs with diverse stuffs and scenes, which are collected from 3D movies, the Internet, and photographs taken by a Fuji W3 stereo camera. **NLPR** contains 1000 image pairs taken by Microsoft Kinect and is often used to evaluate models' ability of multiple salient objects identification. **STEREO** is also called SSB and contains 1000 image pairs collected from 3D movies and the Internet. **LFSD** consists of 100 image pairs, where some images have the characteristic of similar appearance between salient and non-salient regions. **DES** is also called

RGBD135 and consists of 135 image pairs taken by Microsoft Kinect. In our experiments, we regard all image pairs in STEREO, LFSD and DES as testing datasets. Notably, we also select 485 image pairs from NJU2K and 300 image pairs from NLPR to construct two actually used testing datasets, *i.e.* NJU2K-TE and NLPR-TE, respectively.

4.1.2. Evaluation metrics

In this paper, we employ four evaluation metrics S-measure (S) [42], max F-measure (maxF) [43], max E-measure (maxE) [44], and mean absolute error (MAE) [45] to quantitatively evaluate different models' performance.

S-measure measures the structural similarity between the predicted saliency map and the ground truth from region-aware (S_r) and object-aware (S_o), and it is formulated as:

$$S = \alpha * S_o + (1-\alpha) * S_r, \tag{4}$$

where $\alpha \in [0,1]$ denotes the balance parameter. Here, we set α to 0.5 for weighting S_r and S_o in the same extent.

F-measure is a weighted harmonic mean of precision and recall, which is commonly used to evaluate binary classification model, and it is formulated as:

$$F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{5}$$

Notably, to weight precision more than recall, we set β^2 to 0.3 as suggested in [26,43]. Meanwhile, using different thresholds [0, 255], we can obtain the mean F-measure and max F-measure.

E-measure means the enhanced-alignment measure, which considers the local details and global information simultaneously. According to [44], the E-measure is defined as:

$$\begin{aligned} \xi_{SM} &= \frac{2\varphi_{GT} \circ \varphi_{SM}}{\varphi_{GT} \circ \varphi_{GT} + \varphi_{SM} \circ \varphi_{SM}} \\ E &= \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} f(\xi_{SM}) \end{aligned}, \tag{6}$$

where $\circ$ represents the Hadamard product, φ_{GT} and φ_{SM} are two bias metrics for the ground truth and saliency map, and alignment matrix ξ is used to quantify the bias matrix similarity. Besides, $f(\cdot)$ can be defined as a quadratic form function.

MAE provides a fair comparison between the saliency map **SM** and ground truth **GT**, and it is formulated as:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W \times H} |\mathbf{SM}(i) - \mathbf{GT}(i)|, \tag{7}$$

where W and H denote the width and height of the saliency map. Before calculating MAE, we normalize the saliency map and ground truth to [0, 1].

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

In this section, we compare the proposed model with totally fifteen state-of-the-art RGBD saliency models including six traditional models LHM [15], GP [17], LBE [18], SE [29], CDCP [30], MDSF [31] and nine CNNs based models DF [32], CTMF [19], MMCI [22], PCF [20], TANet [23], CPFP [25], DMRA [24], D3Net [34] and AR [26]. Extensive experiments are performed on five RGBD datasets including NJU2K-TE, NLPR-TE, STEREO, LFSD, and DES. Here, we obtain the saliency maps of all fifteen state-of-the-art RGBD saliency models by running source codes or collecting results provided by authors.

As shown in Table 1, the quantitative results comprehensively demonstrate the effectiveness and superiority of the propose model. Concretely, compared with the fifteen state-of-the-art models, our model brings significant performance improvement on three datasets including STEREO, NJU2K-TE and NLPR-TE. For the other two datasets LFSD and DES, the performance of our model is slightly lower than AR on LFSD, and is comparable to AR on DES.

To qualitatively show the excellent performance of our model, visualization comparison results of different RGBD saliency models are given in Fig. 5. Specifically, the low contrast examples shown in the 1st and 2nd rows, where the foreground regions and background regions have similar appearance in RGB images. It can be found that most RGBD saliency models such as AR, D3Net, DMRA, CPFP, LBE can only distinguish the main parts of leaf and sculpture from background, as shown in Fig. 5. Meanwhile, they also highlight some background regions falsely. In contrast, our model achieves the best performance, especially for the 2nd example, where our saliency map not only displays complete salient regions but also suppresses the background regions effectively.

Successively, for the 3rd–5th examples, *i.e.* multiple confused objects, where only one of them is salient object and others are confused objects. Thus, this attribution will cause serious negative impacts for RGBD saliency detection. For example, the 3rd example (the salient object: the Statue of Liberty, the confused objects: chessboard, android and candle), D3Net highlights the regions of the green android robots falsely, DMRA identifies the chessboard as the salient object, both CPFP and MMCI make mistakes on all objects. For the 4th example (the salient object: chocolate, the confused objects: similar patterns on the box), most RGBD saliency models produce very rough saliency maps, where a large number of regions are detected mistakenly. For the 5th example (the salient object: adult, the confused objects: two surrounding children), some models can't completely suppress the two surrounding children such as D3Net, AR, CPFP, TANet, PCF, MMCI, and CTMF. In stark contrast, our model is able to effectively popout the salient object from the confused objects.

The reason behind this can be attributed to the proposed CDFE module and the CFI module. On one hand, the CDFE module can provide multi-level deep depth features, where low-level features $\{\mathbf{F}_i^D\}_{i=1}^3$ contain sufficient spatial contrast information shown in Fig. 2. It can be found that the feature maps of the 3rd example eliminates the confused objects. Meanwhile, the spatial contrast on foreground, background and boundary are clear enough, which is useful for detecting the salient object accurately. On the other hand, by employing six CFI modules in saliency inference branch, we can progressively integrate multi-level cross-modal deep features. During this process, we first carry out the IE step to introduce abundant spatial contrast and depth contexts. Then, the FE step can take the advantage of depth cues to reduce the noises caused by the confused objects. As shown in Fig. 4, the background noises from the confused objects in $\mathbf{FM}_2$ can be filtered effectively. Therefore, our model has enough discrimination to handle such challenging scenes.

Besides, the 6th and 7th examples demonstrate that our model also has the reliable ability to process the scenes with multiple objects. For the 8th example, we can find that the scene not only has a large-scale object, but also is with the similar appearance between the salient object and background. Fortunately, compared with other advanced RGBD saliency models, it is obvious that our model successfully dig the salient objects, which further shows our model's outstanding ability when dealing with such challenging scenes with large-scale salient objects. For the last challenging situation, *i.e.* complex background, as shown in the 9th and 11th rows. We can easily find that saliency maps of some RGBD models, *e.g.* AR, D3Net, DMRA, PCF, are seriously influenced by such complex background. In contrast, saliency maps generated by our model accurately point out the salient objects with clear details, which proves the effectiveness and superiority of the proposed RGBD saliency model again.

Table 1
Overall quantitative comparison among our model and 15 state-of-the-art RGBD saliency models on five RGBD datasets, where four evaluation metrics including S-measure, maxF, maxE and MAE are employed. Meanwhile, the top two scores in each row are highlighted as red and blue, respectively. Notably, ↑ and ↓ indicate that larger and smaller score is better for corresponding metric, respectively.

		LHM	GP	LBE	SE	CDCP	MDSF	DF	CTMF	MMCI	PCF	TANet	CPFP	DMRA	D3Net	AR	Ours
	Metric	[15]	[17]	[18]	[29]	[30]	[31]	[32]	[19]	[22]	[20]	[23]	[25]	[24]	[34]	[26]	
STEREO	S ↑	0.562	0.588	0.660	0.708	0.713	0.728	0.757	0.848	0.873	0.875	0.871	0.879	0.886	0.891	0.893	0.902
	$maxF$ ↑	0.683	0.671	0.633	0.755	0.664	0.719	0.757	0.831	0.863	0.860	0.861	0.874	0.868	0.881	0.886	0.894
	$maxE$ ↑	0.771	0.743	0.787	0.846	0.786	0.809	0.847	0.912	0.927	0.925	0.923	0.925	0.920	0.930	0.930	0.936
	MAE ↓	0.172	0.182	0.250	0.143	0.149	0.176	0.141	0.086	0.068	0.064	0.060	0.051	0.047	0.054	0.053	0.043
NJU2K-TE	S ↑	0.514	0.527	0.695	0.664	0.669	0.748	0.763	0.849	0.858	0.877	0.878	0.879	0.886	0.895	0.893	0.897
	$maxF$ ↑	0.632	0.647	0.748	0.748	0.621	0.775	0.804	0.845	0.852	0.872	0.874	0.877	0.872	0.889	0.891	0.893
	$maxE$ ↑	0.724	0.703	0.803	0.813	0.741	0.838	0.864	0.913	0.915	0.924	0.925	0.926	0.908	0.932	0.930	0.930
	MAE ↓	0.205	0.211	0.153	0.169	0.180	0.157	0.141	0.085	0.079	0.059	0.060	0.053	0.051	0.051	0.055	0.048
LFSD	S ↑	0.553	0.635	0.729	0.692	0.712	0.694	0.783	0.788	0.787	0.786	0.801	0.828	0.847	0.832	0.876	0.862
	$maxF$ ↑	0.708	0.783	0.722	0.786	0.702	0.779	0.813	0.787	0.771	0.775	0.796	0.826	0.849	0.819	0.877	0.866
	$maxE$ ↑	0.763	0.824	0.797	0.832	0.780	0.819	0.857	0.857	0.839	0.827	0.847	0.872	0.899	0.864	0.912	0.891
	MAE ↓	0.218	0.190	0.214	0.174	0.172	0.197	0.145	0.127	0.132	0.119	0.111	0.088	0.075	0.099	0.070	0.074
DES	S ↑	0.578	0.636	0.703	0.741	0.709	0.741	0.752	0.863	0.848	0.842	0.858	0.872	0.901	0.904	0.913	0.912
	$maxF$ ↑	0.511	0.597	0.788	0.741	0.631	0.746	0.766	0.844	0.822	0.804	0.827	0.846	0.857	0.885	0.897	0.903
	$maxE$ ↑	0.653	0.670	0.890	0.856	0.811	0.851	0.870	0.932	0.928	0.893	0.910	0.923	0.945	0.946	0.951	0.945
	MAE ↓	0.114	0.168	0.208	0.090	0.115	0.122	0.093	0.055	0.065	0.049	0.046	0.038	0.029	0.030	0.031	0.027
NLPR-TE	S ↑	0.630	0.654	0.762	0.756	0.727	0.805	0.802	0.860	0.856	0.874	0.886	0.888	0.899	0.906	0.914	0.924
	$maxF$ ↑	0.622	0.611	0.745	0.713	0.645	0.793	0.778	0.825	0.815	0.841	0.863	0.867	0.855	0.885	0.897	0.914
	$maxE$ ↑	0.766	0.723	0.855	0.847	0.820	0.885	0.880	0.929	0.913	0.925	0.941	0.932	0.942	0.946	0.950	0.956
	MAE ↓	0.108	0.146	0.081	0.091	0.112	0.095	0.085	0.056	0.059	0.044	0.041	0.036	0.031	0.034	0.031	0.026

4.3. Ablation studies

To assess the effects of vital components in our model from quantitative and qualitative aspects, we first modify the structure of the blocks utilized by the saliency inference branch to design several variations of our model shown in Fig. 6. Then, we conduct comprehensive experiments on two RGBD datasets NJU2K-TE [16] and LFSD [40]. Correspondingly, the quantitative results are shown in Table 2, and the qualitative results of two challenging situations including low-contrast between foreground and background (the 1st row) and multiple confused objects (the 2nd row) are shown in Fig. 7.

1) **Information Enrichment (IE) and Feature Enhancement (FE) in Cross-modal Feature Integration (CFI) Module:** The CFI module, which mainly contains two steps including information enrichment (IE) and feature enhancement (FE), are employed to insert into the saliency inference branch. To prove the effectiveness of each step in CFI module, we design three variations of the CFI module: (a), "B", *i.e.* the proposed model without depth cues; (b), "B + F(IE)", only adopts information enrichment (IE) step; (c), "B + F(FE)", only adopts feature enhancement (FE) step.
   Specifically, firstly, the quantitative comparison results shown in Table 2 indicate that both the IE and FE steps can exploit the complementary depth information to bring significant performance improvement. In particular, the proposed model, *i.e.* (g) "B + F (IE + FE)", outperforms all variations in terms of S-measure, max F-measure, max E-measure and MAE, and that is to say the best performance can be achieved when the IE and FE steps work together. Then, the qualitative comparison results are shown in Fig. 7. For the 1st example, (a) can't work at all, (b) and (c) mistakenly highlight some background regions, whose depth maps are with high contrast, and in contrast, our model shown in Fig. 7(g) can successfully suppress the background regions and generate a high-quality saliency map. As for the 2nd example, only (g) doesn't classify the surround persons as salient objects mistakenly. Therefore, we can make a conclusion that jointly performing the IE and FE steps can sufficiently integrate the cross-modal features and endow the saliency model with more powerful discrimination ability.
2) **Multi-level Deep Depth Features:** In our model, the available depth information contains original depth maps, enhanced depth maps and multi-level deep depth features. In [25], they choose to exploit enhanced depth maps and use them to enhance the RGB features. Differently, our model employs multi-level deep depth features in the saliency inference branch. Thus, to explore the rationality of using multi-level deep depth features, we also design some variations of our model: "B + D(IE + FE)" shown in Fig. 6(d) uses original depth maps, and "B + ED(IE + FE)" shown in Fig. 6(e) uses enhanced depth maps. Notably, to make sure the depth map's resolution and channel number match with multi-level depth features, we first repeat original depth maps or enhanced depth maps to construct six blocks, where each block's channel number is 64, and then we perform bilinear downsampling on these depth maps.

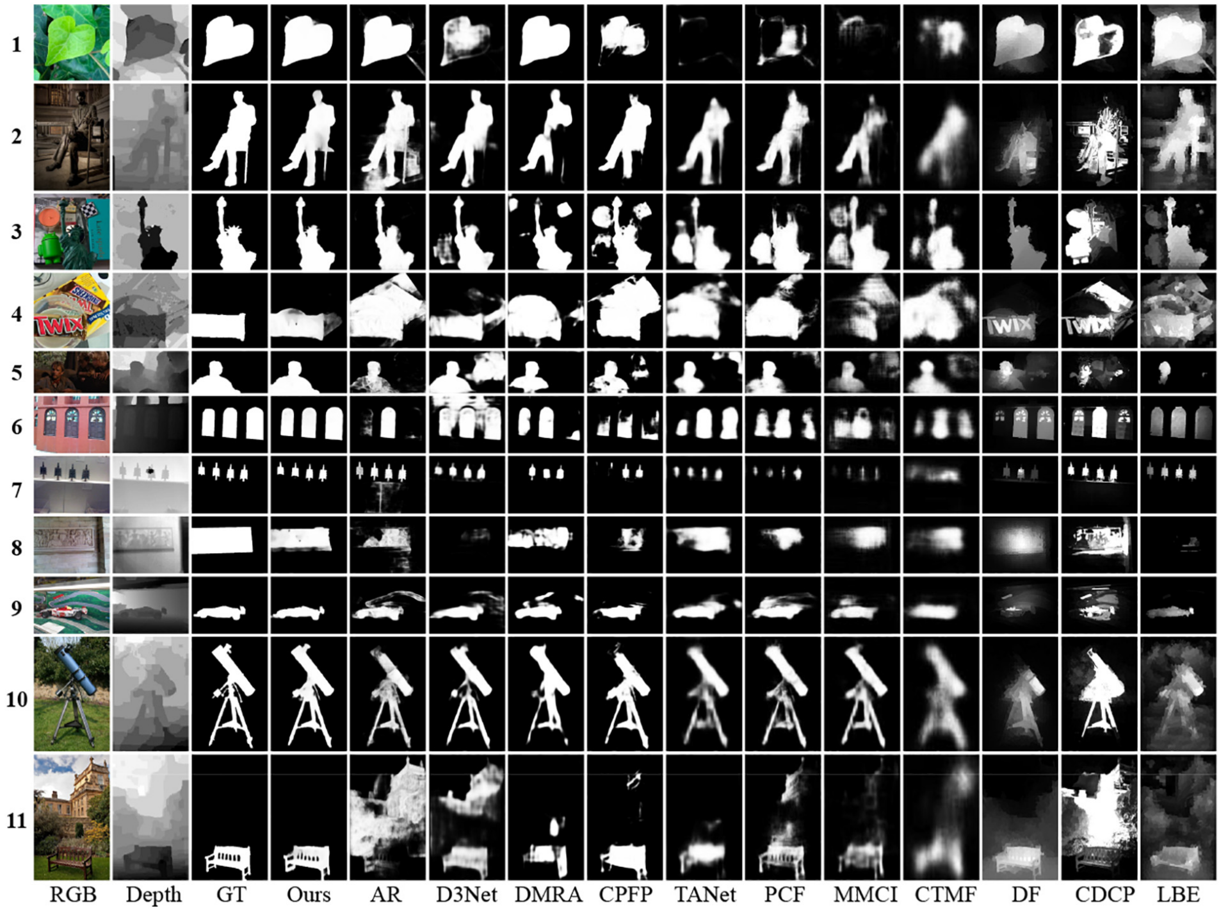
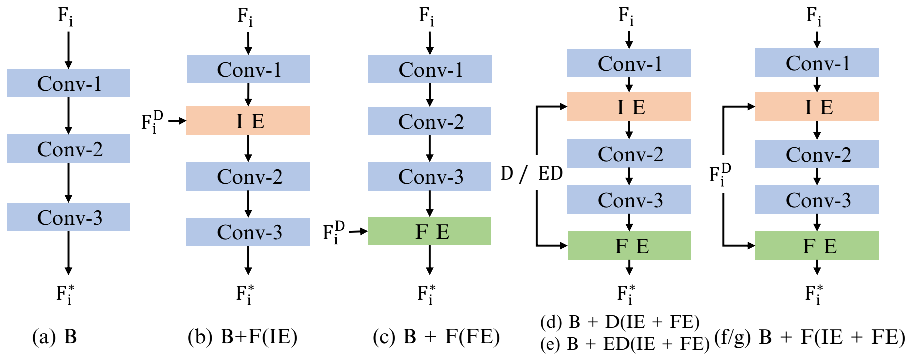


Fig. 5. Visualization comparisons over the proposed model and several state-of-the-art RGBD saliency models including nine CNNs based models and two traditional models.



(a) B (b) B+F(IE) (c) B + F(FE) (d) B + D(IE + FE) (e) B + ED(IE + FE) (f/g) B + F(IE + FE)

Fig. 6. Illustration of several variations of the proposed RGBD saliency model. Notably, in (d) and (e), “D” denotes depth maps, “ED” represents enhanced depth maps.

Table 2
Quantitative results of variations of the proposed model on two RGBD datasets including NJU2K-TE [16] and LFSD [40]. Top two scores in each row are highlighted as red and blue, respectively.

		(a) B	(b) B+F(IE)	(c) B+F(FE)	(d) B+D(IE+FE)	(e) B+ED(IE+FE)	(f) B+F(IE+FE)*	(g) B+F(IE+FE)
NJU2K-TE	$S \uparrow$	0.873	0.888	0.886	0.895	0.893	0.893	0.897
	$maxF \uparrow$	0.865	0.885	0.882	0.890	0.892	0.890	0.893
	$maxE \uparrow$	0.912	0.921	0.920	0.925	0.928	0.926	0.930
	$MAE \downarrow$	0.060	0.054	0.055	0.050	0.050	0.051	0.048
LFSD	$S \uparrow$	0.811	0.841	0.837	0.841	0.863	0.853	0.862
	$maxF \uparrow$	0.821	0.841	0.851	0.852	0.855	0.853	0.866
	$maxE \uparrow$	0.853	0.881	0.883	0.875	0.892	0.887	0.891
	$MAE \downarrow$	0.108	0.088	0.094	0.088	0.077	0.083	0.074

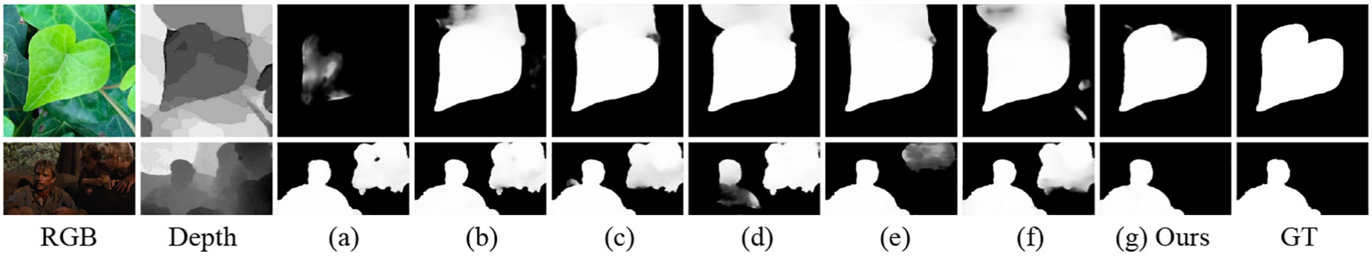


Fig. 7. Visualization comparisons of different variations. (a): B, (b): B + F(IE), (c): B + F(FE), (d): B + D(IE + FE), (e): B + ED(IE + FE), (f): B + F(IE + FE)*, (g): B + F(IE + FE).

Concretely, firstly, the quantitative comparison results are shown in Table 2, which indicates three facts: (I), by comparing (d, e) with (a), we can demonstrate the superiority of the CFI module; (II), by comparing (d) with (e), we can verify that depth contrast is useful for the quality improvement of depth maps; (III), by comparing (d), (e) with (g), we can firmly demonstrate the effectiveness of multi-level deep depth features, and it also demonstrates the rationality of our model's design. Besides, for the 2nd example in the qualitative comparison results shown in Fig. 7, the depth distribution of surrounding persons is more significant than the salient objects. It can be seen that (d) produces a complete different result compared with the ground truth, (e) can identify salient objects correctly, but its results still falsely highlight some background regions. In contrast, our model shown in Fig. 7(g) acquires the best saliency map compared with the ground truth. Overall, we can make a conclusion that the multi-level deep depth features generated by CDFE module are more effective than the original depth maps or the enhanced depth maps.

3) **Depth Contrast based Deep Depth Features:** The multi-level deep depth features are extracted by the CDFE module, which is trained with the contrast loss. To measure the contribution of the contrast loss in our model, we define a variation of our model, named as (f) “B + F(IE + FE)*”, which replaces the contrast loss with a conventional saliency loss employed by the CFI module. Specifically, (f) attempts to force the CDFE module to directly detect salient objects from depth maps, while our model attempts to increases the depth contrast between foreground and background. The corresponding comparison results are shown in Table 2, *i.e.* the quantitative comparison results, and Fig. 4, *i.e.* the qualitative comparison results. According the two comparison results, it can be found that the performance of employing the contrast loss, *i.e.* our model, is better than the model using the saliency loss (f). Therefore, this clearly demonstrates the effectiveness of contrast loss in our model and also shows the rationality of the employment of contrast loss.

5. Conclusion

This paper proposes a novel RGBD saliency model, which consists of the RGB branch, the depth branch and the saliency inference branch, to effectively pop-out salient objects. The key components in our model focus on the contrast-guided depth feature extraction (CDFE) module and the cross-modal feature integration (CFI) module. Specifically, the CDFE module, namely the depth branch, is designed to provide multi-level deep depth features, which not only contain spatial details but also supply rich context cues. Successively, the CFI module involving information enrichment step and feature enhancement step, is embedded into the saliency inference branch. Thus, the cross-modal cues including the multi-level deep RGB and depth features are progressively integrated into the final high-quality RGBD saliency map. Comprehensive experiments are performed on five public RGBD datasets, and the experimental results clearly demonstrate the effectiveness and superiority of our model.

CRediT authorship contribution statement

Liang Pan: Methodology, Software, Writing - original draft. **Xiaofei Zhou:** Conceptualization, Methodology, Validation, Writing - review & editing. **Ran Shi:** Conceptualization, Writing - review & editing. **Jiyong Zhang:** Writing - review & editing, Supervision. **Chenggang Yan:** Funding acquisition, Project administration.

Declaration of competing interest

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contribution to this work.

We confirm that the manuscript has been read and approved by all authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the corresponding author is the sole contact for the Editorial process.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61901145, Grant 61801219 and Grant 61931008.

References

[1] P. Mukherjee, B. Lall, Saliency and kaze features assisted object segmentation, J. Image Vision Comput. 61 (2017) 82–97.
[2] Z. Li, S. Qin, L. Itti, Visual attention guided bit allocation in video compression, J. Image Vision Comput. 29 (2011) 1–14.
[3] R. Quispe, H. Pedrini, Improved person re-identification based on saliency and semantic parsing with deep neural network models, J. Image Vision Comput. 92 (2019) 103809.
[4] C. Liu, P.C. Yuen, Human action recognition using boosted eigenactions, J. Image Vision Comput. 28 (2010) 825–835.
[5] Y. Chuan, Z. Lihe, L. Huchuan, R. Xiang, Y. Ming-Hsuan, Saliency detection via graph-based manifold ranking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE 2013, pp. 3166–3173.
[6] W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE 2014, pp. 2814–2821.
[7] 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.
[8] 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.
[9] 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.

[10] 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.

[11] 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.

[12] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, Basnet: boundary-aware salient object detection, The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[13] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, International Conference on Computer Vision, ICCV, IEEE 2015, pp. 2650–2658.

[14] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab, Deeper depth prediction with fully convolutional residual networks, International Conference on 3D Vision, 3DV, IEEE, 2016.

[15] 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.

[16] 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.

[17] J. Ren, X. Gong, L. Yu, W. Zhou, M.Y. Yang, Exploiting global priors for rgb-d saliency detection, Computer Vision and Pattern Recognition Workshops, CVPRW, IEEE 2015, pp. 25–32.

[18] 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.

[19] 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.

[20] 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.

[21] H. Chen, Y.-F. Li, D. Su, Attention-aware cross-modal cross-level fusion network for rgb-d salient object detection, International Conference on Intelligent Robots and Systems, IROS, IEEE 2018, pp. 6821–6826.

[22] 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 Recogn. 86 (2019) 376–385.

[23] H. Chen, Y. Li, Three-stream attention-aware network for rgb-d salient object detection, IEEE Trans. Image Process. 28 (6) (2019) 2825–2835.

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

[25] 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.

[26] X. Zhou, G. Li, C. Gong, Z. Liu, J. Zhang, Attention-guided rgbd saliency detection using appearance information, J. Image Vision Comput. 95 (2020) 103888.

[27] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representations 2014, pp. 1–14.

[28] K. He, X. Zhang, S. Ren, S. Jian, Deep residual learning for image recognition, Computer Vision and Pattern Recognition, CVPR, IEEE 2016, pp. 770–778.

[29] 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.

[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] 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.

[32] 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.

[33] 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.

[34] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, M.-M. Cheng, Rethinking rgb-d salient object detection: models, datasets, and large-scale benchmarks, arXiv preprint, arXiv:1907.06781.

[35] S. Xie, Z. Tu, Holistically-nested edge detection, International Conference on Computer Vision, ICCV, IEEE 2015, pp. 1395–1403.

[36] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, A.Desmaison Lin L., A. Antiga, Lerer, Automatic differentiation in pytorch, 2017.

[37] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, vol. 9, 2010, pp. 249–256.

[38] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, International Conference on Learning Representations, ICLR, 2015.

[39] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, Computer Vision and Pattern Recognition, CVPR, IEEE 2012, pp. 454–461.

[40] 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.

[41] 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.

[42] 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.

[43] R. Achanta, S.S. Hemami, F.J. Estrada, S. Süsstrunk, Frequency-tuned salient region detection, Computer Vision and Pattern Recognition, CVPR, IEEE 2009, pp. 1597–1604.

[44] 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 27th International Joint Conference on Artificial Intelligence 2018, pp. 698–704.

[45] A. Borji, D.N. Sihite, L. Itti, Salient object detection: a benchmark, European Conference on Computer Vision, ECCV, Springer 2012, pp. 414–429.