Neurocomputing 451 (2021) 95–106

Contents lists available at ScienceDirect

Neurocomputing

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

Circular Complement Network for RGB-D Salient Object Detection

Zhen Bai [a,b], Zhi Liu [a,b,*], Gongyang Li [a,b], Linwei Ye [c], Yang Wang [d]

[a] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China
[b] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
[c] College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China
[d] Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

ARTICLE INFO

Article history:
Received 21 June 2020
Revised 12 March 2021
Accepted 14 April 2021
Available online 1 May 2021
Communicated by Zidong Wang

Keywords:
RGB-D salient object detection
Depth cues
Circular feature complement
Transmission Bridge

ABSTRACT

With the supplement of texture and geometry cues in depth maps, some difficult scenes of salient object detection (SOD) in 2D images can be overcome. However, some distractors in the depth maps with relatively poor quality may interfere with SOD. Thus, how to suppress the interference of depth maps and extract valuable depth cues, is a critical issue to serve as effective complements to RGB cues. Aiming at addressing this issue, we propose a predict-refine scheme based Circular Complement Network (CCNet), which consists of a prediction subnetwork and a refinement subnetwork. On one hand, since RGB images generally contain more essential information for SOD, we propose a strategy which employs higher-level RGB feature maps to suppress the interference of depth feature maps. With this strategy, a novel Circular Feature Complement (CFC) module is specifically designed to enhance depth feature maps as well as to promote mutual complementarity between RGB feature maps and depth feature maps. The CFC modules are embedded into two subnetworks to achieve the cross-modal interactions at three levels. On the other hand, for the sake of the integration of two subnetworks, a Transmission Bridge (TB) module is proposed to effectively transfer the feature maps of the prediction subnetwork to the refinement subnetwork. The non-salient regions are thus further suppressed in the TB module. Comprehensive experiments on six benchmark datasets show that the proposed CCNet outperforms 13 state-of-the-art models.

1. Introduction

Salient object detection (SOD) is stimulated by human visual attention mechanism to highlight the most attractive and distinctive objects/regions in a scene. It has been successfully applied in a variety of applications such as image retrieval [1], image compression [2] and image segmentation [3–7]. With the development of distance sensing technology in the field of computer vision, 3D data dedicated to depicting 3D scenes of real life fills our lives. With this situation, SOD has also extended from 2D images to RGB-D images. The RGB SOD task [8,9] has come of age with the earlier lots of research attentions. However, due to low contrasts and disturbing background regions in some complex RGB images, the detection quality of RGB SOD for some intricate scenes degrades significantly. Compared with RGB images, depth maps are not easy to be interfered by light changes [10–12]. Thus, adding depth maps can improve the structural integrity of salient objects with the complements of geometry and texture cues of objects [13–15], which have been consistently demonstrated in some studies [10,14,16].

RGB images and depth maps belong to different modalities, thus in the past decade, a large number of models [10–12,15,17–32] focused on exploring the cross-modal complementarity of RGB images and depth maps. With the success of convolutional neural network (CNN) in learning discriminative features [33], most of RGB-D SOD models proposed in recent years are based on CNN [20–30,34]. These models generally followed a two-stream architecture [35,36] and combined RGB feature maps with depth feature maps at multiple levels to improve the saliency detection performance.

However, these models are not robust for the scenes with relatively poor-quality depth maps, as shown in the results of DMRA [20] in Fig. 1, the generated saliency maps are contaminated as a result of some interferences (shown in red bounding-boxes) of poor-quality depth maps. The two-stream models usually directly introduced the depth maps/feature maps without any effective enhancement processes [20–30], which leads to that the interference of depth maps/feature maps can not be effectively suppressed in the cross-modal complementary processes. Zhao *et al.* [29] tried to employ contrast-enhanced network to enhance the contrast of

* Corresponding author at: Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China.
E-mail address: liuzhisjtu@163.com (Z. Liu).

https://doi.org/10.1016/j.neucom.2021.04.052
0925-2312/

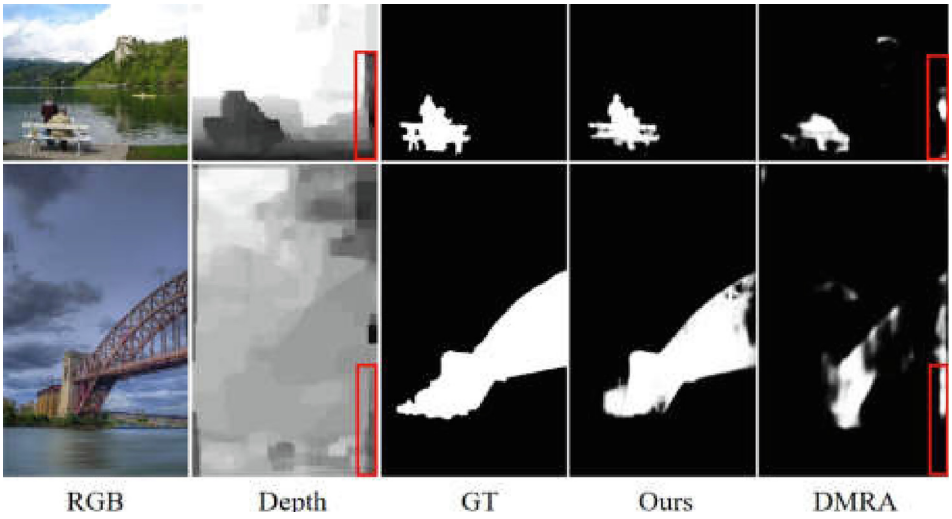


Fig. 1. Some results of the proposed model and one state-of-the-art model.

depth map, but there is no direct mapping relationship between depth map and saliency map. This network can only work for the depth map where salient regions obviously discriminate from non-salient regions, but it fails to act on the depth map with low contrasts. To prevent contamination from unreliable depth information, Fan *et al.* [37] designed a model which discards the poor-quality depth maps automatically through a depth depurator unit. However, removing the poor-quality depth map indicates that the discriminative depth information for SOD is also discarded. Therefore, how to extract valuable depth features to complement RGB features more effectively, even from poor-quality depth maps, is still a challenging issue to be solved.

To solve this issue, we propose a predict-refine scheme based Circular Complement Network (CCNet), which mainly consists of a prediction subnetwork and a refinement subnetwork, with the proposed Circular Feature Complement (CFC) module and Transmission Bridge (TB) module. The CFC module integrates two-modal feature maps in a circular complement manner, *i.e.* RGB-to-depth direction and depth-to-RGB direction. These two opposite directions of integration work together for the mutually cross-modal complementarity. To promote the integrity of CCNet, the TB module is proposed as a bridge to transmit the output of the prediction subnetwork to the refinement subnetwork and to narrow the gap between them.

The CFC module is the key to address the challenging issue of cross-modal complementarity. RGB images differentiate from depth maps as they are more informative and generally contribute more to SOD. Inspired by this observation, we propose a strategy that utilizes the RGB feature maps to enhance depth feature maps for integration in the RGB-to-depth direction. Meanwhile, considering that the effectiveness of this strategy relies on the representation ability of RGB features and the CNN features at deeper layers are more representative, the relatively higher-level RGB feature maps are employed to suppress the interference of non-salient regions of relatively lower-level depth feature maps in the RGB-to-depth direction. This boosts the representation ability of depth feature maps. For the depth-to-RGB direction, with the supplement of the boosted depth feature maps, the RGB feature maps can be exploited to successfully highlight salient objects, benefiting from salient object continuity in cross-modal features. In TB module, the preliminary saliency map generated from the prediction subnetwork makes available for focusing on the salient regions in the middle-level feature maps extracted from the prediction subnetwork. In this way, more valuable RGB and depth information are leveraged by the refinement subnetwork to better locate salient objects and sharpen the preliminary results. Extensive experiments demonstrate that the proposed CCNet outperforms favorably against 13 state-of-the-art models.

Our main contributions can be summarized as follows:

1) We propose a predict-refine scheme based Circular Complement Network (CCNet), which is equipped with the CFC module and the TB module. Our CCNet can progressively suppress the interference of poor-quality depth maps and effectively explore cross-modal complementarity to enhance the integrity of salient objects.
2) We propose the CFC module, which first utilizes RGB features to enhance depth feature maps (RGB-to-depth), and then adopts the enhanced depth feature maps in turn to assist RGB feature maps to highlight salient objects (depth-to-RGB). The two opposite directions of feature integration achieve the cross-modal complementarity of depth and RGB feature maps in a circular complement manner.
3) We propose the TB module to effectively connect the prediction subnetwork and the refinement subnetwork. It takes the preliminary saliency map of the prediction subnetwork to focus on the salient regions in feature maps. The transferred feature maps of TB module are crucial for the completeness of salient objects in the refinement subnetwork.

2. Related work

With the increasing research attention in the field of RGB-D SOD, more and more models have been proposed. In this section, we classify existing RGB-D SOD models into traditional RGB-D SOD models and CNN-based RGB-D SOD models, and briefly review them. Besides, we also review some landmark models of visual attention prediction and image salient object detection tasks.

2.1. Traditional RGB-D SOD models

Lang *et al.* [16] and Niu *et al.* [10] made preliminary studies for RGB-D SOD, they introduced depth map to SOD and proved the validity and necessity of depth map. Subsequently, based on multi-contextual depth contrast, Peng *et al.* [13] extended RGB SOD models by incorporating depth-induced saliency results. Cheng *et al.* [38] and Fan *et al.* [17] combined depth contrast feature with color and spatial features to detect salient object in

RGB-D scenes. The above-mentioned Models fused RGB-induced saliency map and depth-induced saliency map, but they lack cross-modal complementarity. According to some observations, visual attention of human is more easily to be paid to regions closer to them and several models have been proposed. Ju *et al.* [39] took object-to-surrounding depth contrast to detect salient objects. Ren *et al.* [40] depicted depth prior at first, and then fused the depth prior, global-context surface orientation prior, background prior and region contrast to guide SOD. Feng *et al.* [19] constructed depth saliency feature based on the angular density and size in depth distributions to quantify saliency. Liang *et al.* [41] integrated color contrast, disparity contrast and depth-guided-background priors to detect saliency of 3D stereoscopic scenes. Though the above models [19,39–41] have promoted the progress of traditional models, they were limited by the previous observations and were hard to generalize to all scenes. Song *et al.* [31] measured saliency on four classes of features at three levels, and took random forest regression to fuse saliency at multiple scales. Wang *et al.* [42] proposed a multistage-based model by using the minimum barrier distance transformation and multilayer cellular automata-based saliency fusion. Wang *et al.* [43] focused on addressing the stereoscopic image SOD, which is similar to the RGB-D SOD, and they employed the disparity of pairs of images taken from different angles to detect salient objects in stereo scenes. In summary, the performance of these traditional RGB-D SOD models is limited by the hand-crafted feature.

2.2. CNN-based RGB-D SOD models

The fast development of deep learning and the proliferation of RGB-D labeled data greatly promote the study of RGB-D SOD. Numerous CNN-based RGB-D SOD models have been proposed in recent years. At first, Qu *et al.* [32] extracted the hand-craft feature from the RGB images and paired depth maps, and then took these features as the input of a CNN. Finally, the outputted classification probability values were defined as the saliency values in region level. Han *et al.* [22] constructed a two-stream framework which lately fuses CNN features of RGB and depth to predict saliency map. Since then, the two-stream framework has been the mainstream in the study of CNN-based RGB-D SOD [20–24,28–30,44]. Zhu *et al.* [30] extracted depth feature from a designed subnetwork, and combined depth feature with RGB feature in a middle stage through cascade. The result fusion strategy was employed in [18], Wang *et al.* designed an adaptive fusion module to adaptively fuse the saliency maps of two streams. Considering the above models lack the effect use of complementary information of the two modalities, Chen *et al.* [23] proposed cross-modal complementarity modules to fuse two-modal features at multiple levels to progressively enhance saliency predictions. Whereafter, the cross-modal complement modules with complementarity-aware supervisions are explored by Chen *et al.* [21] across all levels. Chen *et al.* [24] later integrated attention-aware cross-level fusion modules to select discriminative feature from two modalities. Piao *et al.* [20] combined depth cues with multi-scale context features, and designed a recurrent attention module to boost the performance of SOD. Li *et al.* [28] employed discriminative operations for features from RGB images and depth maps and achieved cross-modal complementary at each level. Along with the thorough research for RGB-D saliency detection, it is found that the poor-quality depth map imposes negative impacts on RGB-D SOD, and this situation has been taken account into some models [29,44]. Zhou *et al.* [44] extracted paired modal features from RGB stream and depth stream, and fused them under the guidance of contrast-enhanced depth maps at multiple levels, Zhao *et al.* [29] designed a fluid pyramid integration framework which took enhanced depth map to enhance RGB features at multiple levels. However, the contrast-enhanced strategy adopted for depth map/feature lacks generality in the real scenes. In addition to the two-stream based models, Chen *et al.* [12] expanded two-stream architecture to three-stream architecture by integrating a cross-modal distillation stream which accompanies the RGB and depth streams.

The proposed CCNet inherits the two-stream style and further detects more accurate salient objects with the predict-refine scheme. Moreover, the CFC module employs RGB feature maps to enhance depth feature maps before achieving cross-modal complementarity between them. Equipped with the CFC module, the proposed CCNet is robust when confronting poor-quality depth maps.

2.3. Visual attention prediction models

Visual attention prediction (*i.e.*, fixation prediction) simulates attentional capability of human perception to predict scene locations where humans may fixate at first glance. Most traditional visual attention prediction models are based on the bottom-up mechanism. Classically, Itti *et al.* [45] combined the contrasts in intensity, color, and orientation to identify the fixation point. Judd *et al.* [46] proposed a bottom-up and top-down visual attention prediction model based on low, middle and high-level image features. Based on decision theory, Ngo *et al.* [47] developed the multi-scale discriminant saliency technique for visual attention prediction. Harel *et al.* [48] introduced the graphical model into visual attention prediction. More details could be found in [49], which is a comprehensive summary of traditional models.

With the promotion of CNN in many tasks of computer visual [50], Vig *et al.* [51] made the first attempt to leverage CNNs for visual attention prediction in an end-to-end manner. They first extracted representative features from image, and then fed them into a classifier for visual attention prediction. Wang *et al.* [52] incorporated multi-scale features trained in a multi-scale supervised manner to infer attention. These models effectively formulated saliency as a regression problem, and outperformed the traditional attention prediction models. Based on the encoder-decoder architecture, Wang *et al.* [53] proposed a pithy method to sense the local and global features for visual attention prediction. More works of CNN-based visual attention prediction could be found in the recent review [54].

2.4. Image salient object detection models

Image salient object detection aims at segmenting out the most attractive objects from an RGB image. Similar to visual attention prediction, most of the traditional image SOD models are based on hand-crafted features and classic technologies, such as cognitive assumptions [55], the over segmentation method [56], superpixels [57], object proposals [6], graph cuts [58] and random walks [59]. Shen *et al.* [60] proposed a higher order binary energy function to achieve binary image segmentation. In [61], Shen *et al.* solved the SOD task through maximizing the proposed knapsack constrained submodular energy function. Wang *et al.* [62] took superpixels as computing units, and transferred the saliency scores from labeled images onto the detected images through matching the similarity of labeled image and detected images. Ren *et al.* [63] promoted the image SOD performance based on similar images. More traditional image SOD models could be found in [64].

CNN-based image SOD models refresh the previous records. Li *et al.* [65] aggregated multi-scale features to infer saliency. Wang *et al.* [66] captured fixation map from high-level CNN features, and then progressively segmented out salient objects with the guidance of the fixation map in a top-down manner. Wang *et al.* [67] proposed a top-down and bottom-up saliency inference model in a joint and iterative way. In [68], to obtain complete salient objects with more precise object boundaries, Zhao *et al.* [68] pro-

posed the two-branch network to generate the boundary map and the preliminary saliency map, respectively, and then they combined these two maps to generate the final saliency map. Qin *et al.* [69] attempted to add boundary loss into an U-Net based model to boost the ability of boundary capture. Wang *et al.* [70] specifically designed a salient edge module to refine the boundary of detected salient objects. In [8], Huang *et al.* proposed a multi-level integration and multi-scale fusion neural network. Wu *et al.* [9] introduced the adversarial learning into image SOD. At present, a plenty of image SOD models reached promising performance in most of scenes. More works of CNN-based image SOD could be found in the recent survey [71].

3. The proposed approach

This section starts with the overall architecture of the proposed CCNet in Section 3.1. Then, the description of the CFC module is presented in Section 3.2, and the detailed formula of the TB module is given in Section 3.3. Finally, the implementation details of CCNet are presented in Section 3.4.

3.1. Overall architecture

As shown in Fig. 2, with three CFC modules and one TB module, the prediction subnetwork and the refinement subnetwork are combined to form the CCNet. The adopted predict-refine scheme in CCNet is generally used to promote the performance of models [69,72], which is well-tried for RGB image SOD. Based on the predict-refine scheme, we extract RGB feature maps and depth feature maps with two streams and employ VGG16 [73] as the backbone of CCNet.

The prediction subnetwork is in charge of predicting the preliminary saliency map from the input RGB image I and the paired depth map D. In this subnetwork, the RGB stream adopts the first five convolution blocks of VGG16 to extract RGB feature maps $\{\mathbf{f}_{i,j}, i = 1, j = 1,2,3,4,5\}$, and the depth stream abstracts depth feature maps $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\}$ via three convolution blocks of VGG16. We feed $\{\mathbf{f}_{1,2}, \mathbf{d}_1\}$ and $\{\mathbf{f}_{1,3}, \mathbf{d}_2\}$ into two CFC modules to achieve the cross-modal fusion of RGB and depth feature maps in the prediction subnetwork at low- and middle-levels, and renamed these CFC modules as CFC-L and CFC-M, respectively. Afterwards, a cascaded partial decoder [74] subsequently integrates three-layer RGB feature maps into the RGB stream to output the preliminary saliency map $\mathbf{S}_{\text{pre}}$ with the size of 88×88.

The prediction subnetwork provides the basis for the following refinement subnetwork. The TB module acts as a bridge between the prediction subnetwork and refinement subnetwork, which utilizes $\mathbf{S}_{\text{pre}}$ to boost middle-level feature maps and transmits them to the refinement subnetwork. The refinement subnetwork subsequently makes further abstractions for imported feature maps, boosting the performance of CCNet. The fourth and fifth convolution blocks of VGG16 are employed in the RGB stream to extract high-level RGB feature maps $\{\mathbf{f}_{i,j}, i = 2, j = 4,5\}$. In the depth stream, another fourth convolution block of VGG16 is used to extract high-level depth feature maps $\mathbf{d}_4$. Through CFC-H module, we achieve the cross-modal feature complementarity at high level in the refinement subnetwork. The decoding operation in the refinement subnetwork is performed in the same way as that in the prediction subnetwork, and the final saliency map $\mathbf{S}_{\text{final}}$ with the size of 88×88 is generated.

More recently, the commonly used models are multi-scale prediction models [28], which are generally constructed with multiple prediction subnetworks. In multi-scale prediction models, the same scale features of decoder and encoder are concatenated, and then they are fed into the corresponding subnetwork to predict saliency map. The final result is generated by the fusion of multiple predictions or gradually predicted by employing the multi-scale features in a coarse-to-fine fashion. The proposed CCNet, which adopts a predict-refine scheme, is essentially different from multi-scale prediction models. The two decoders in the prediction subnetwork and the refinement subnetwork all integrate middle- and high-level RGB features for detection. For subsequent refinement, the prediction subnetwork plays the role of preliminary prediction and extraction of middle-level feature. After the process of purifying middle-level features in the TB module, the enhanced features flow into the refinement subnetwork. The refinement subnetwork makes further feature extractions, and achieves mutual cross-modal complementarity again at high level. With the purified middle-level features and further complemented high-level features, the final results generated by the refinement subnetwork are finer than the preliminary results.

3.2. Circular Feature Complement module

How to extract valuable depth feature maps to complement RGB feature maps effectively, even from poor-quality depth maps, is the key focus of this paper. Aiming at solving this challenging issue, we specifically design a CFC module, as shown in Fig. 3.

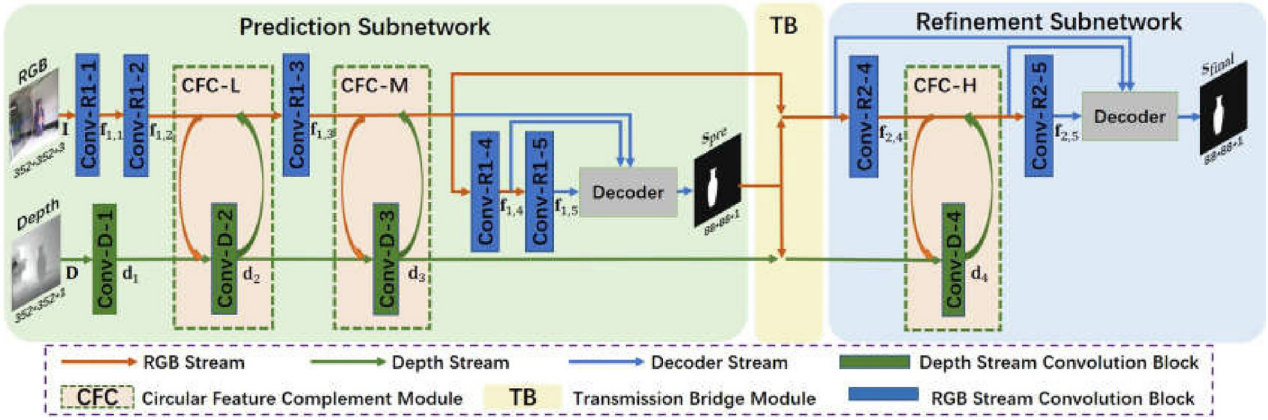


Fig. 2. The overall architecture of the proposed CCNet. The proposed CCNet is constructed based on the predict-refine scheme with two streams. In the prediction subnetwork, a RGB image and a paired depth map are passed over multiple convolution blocks, CFC-L and CFC-M modules and a decoder for preliminary prediction. RGB features and depth features are mutually complemented in CFC modules. The TB module processes the outputs of prediction subnetwork and transfers them to the refinement subnetwork. The refinement subnetwork employs convolution blocks and CFC-H module at high level to generate final saliency map through another decoder. CFC-L, CFC-M, and CFC-L have the same structure, and work for low-, middle-, and high-level features, respectively.

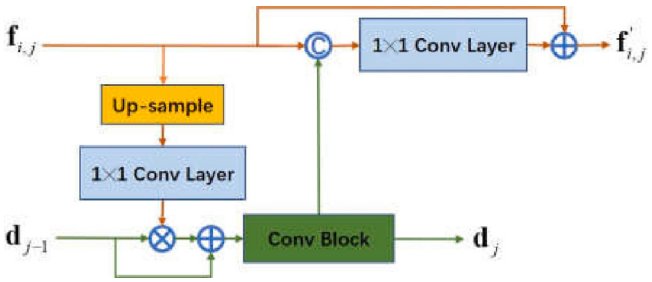


Fig. 3. The structure of CFC module.

Before transferring depth information to RGB stream through the general depth-to-RGB [23,29,30] integration, our CFC module adopts an opposite direction integration which is RGB-to-depth. The RGB-to-depth integration boosts the quality of depth feature maps, and then it works with depth-to-RGB integration to achieve the mutual cross-modal complementarity in a circular complement manner.

The detailed formula of RGB-to-depth integration is designed according to the differences between RGB images and depth maps. Depth maps differentiate from RGB images as they provide more meaningful information about the shape and distance of objects, while RGB images usually focus more on the textures and colors to stimulate eyes. Thus, the mapping relationship between RGB images and saliency maps is relatively clearer than that between depth maps and saliency maps. Based on this observation, we propose a strategy that utilizes relatively discriminative RGB feature maps $\mathbf{f}_{i,j}$ to enhance depth feature maps $\mathbf{d}_{j-1}$ via an element-wise multiplication. We preserve the original depth information by piling the $\mathbf{d}_{j-1}$ onto the initial enhanced depth features by a residual connection, *i.e.* element-wise summation. Considering that the effectiveness of this strategy relies on the representation ability of RGB features and the CNN features at deeper layers are more representative, we employ the RGB feature maps $\mathbf{f}_{i,j}$, which are relatively deeper than the depth feature maps, to enhance depth feature maps $\mathbf{d}_{j-1}$. The RGB-to-depth integration in the CFC module is formulated as follows:

$$\mathbf{d}_j = Conv(\mathbf{d}_{j-1} \otimes conv(up(\mathbf{f}_{i,j})) \oplus \mathbf{d}_{j-1}), \tag{1}$$

where $up(\cdot)$, $conv(\cdot)$, and $Conv(\cdot)$ are up-sampling operation, 1×1 convolution layer and VGG convolution block, respectively. $\otimes$ and $\oplus$ denote element-wise multiplication and element-wise addition, respectively. We adopt up-sampling operation and 1×1 convolution operation to resize $\mathbf{f}_{i,j}$ to fit the resolution of $\mathbf{d}_{j-1}$. For the depth-to-RGB integration, enhanced $\mathbf{d}_j$ are transferred to RGB stream by concatenating with $\mathbf{f}_{i,j}$, then $\mathbf{f}_{i,j}$ is piled onto the composite features via element-wise summation to generate $\mathbf{f}'_{i,j}$. The depth-to-RGB integration is expressed as follows:

$$\mathbf{f}'_{i,j} = conv(Cat(\mathbf{f}_{i,j}, \mathbf{d}_j)) \oplus \mathbf{f}_{i,j}, \tag{2}$$

where $Cat(\cdot,\cdot)$ represents cross-channel concatenation operation. The 1×1 convolution layer is used to compress the channel number of the concatenated feature maps to match $\mathbf{f}_{i,j}$.

In Fig. 4, we show the visualized features in CFC-M module to verify the rationality of RGB-to-depth and depth-to-RGB integrations. Concretely, $\mathbf{f}_{1,3}$ enhances $\mathbf{d}_2$ in the RGB-to-depth integration to generate $\mathbf{d}_3$, and $\mathbf{d}_3$ reversely supplements $\mathbf{f}_{1,3}$ in the depth-to-RGB integration to generate $\mathbf{f}'_{1,3}$. The salient objects in $\mathbf{f}_{1,3}$ are confused with non-salient regions, and some interferences exist in $\mathbf{d}_2$. After two integrations, the interferences are suppressed in $\mathbf{d}_3$, and the salient objects are effectively highlighted in $\mathbf{f}'_{1,3}$. The examples in Fig. 4 demonstrate that the two integrations of CFC are reasonable and effective. With the two integrations combining in a circular complement manner, the CFC module can purify the interferences of depth map and improve the discriminability of features to promote the accuracy of salient object prediction (see Fig. 4).

Through stacking three CFC modules into CCNet, the cross-modal complementarity of RGB and depth feature maps is achieved at three different levels with weakening the contamination from the poor-quality depth maps, which progressively promotes SOD.

3.3. Transmission Bridge module

As described in Section 3.1, on the basis of the adopted predict-refine scheme in CCNet, the effectiveness of the refinement subnetwork is prone to be affected by the representation ability of inputted feature maps. To this end, a TB module is added before the refinement subnetwork to enhance the discrimination of imported feature maps. Meanwhile, the TB module can serve as a bridge of the prediction subnetwork and the refinement subnetwork to improve the integrity of the CCNet.

Although the preliminary predictions can provide localization cues to suppress the background regions in original RGB and depth feature maps, they are relatively coarse and some boundary information of salient objects may be filtered out. This situation will wrongly suppress the boundary features in the boosted feature maps. To further purify the preliminary predictions, we introduce

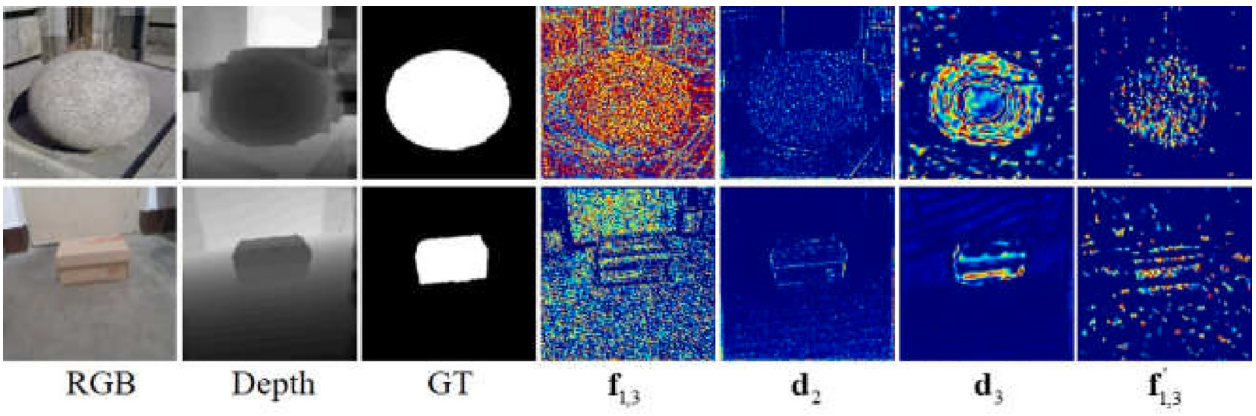


Fig. 4. Feature visualization in CFC-M. $\mathbf{f}_{1,3}$ and $\mathbf{d}_2$ are the input RGB feature and depth feature of CFC-M, respectively. $\mathbf{d}_3$ and $\mathbf{f}'_{1,3}$ are the output RGB feature and depth feature of CFC-M, respectively.

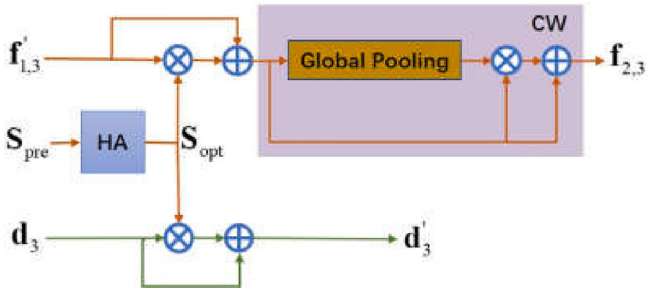


Fig. 5. The structure of TB module.

the Holistic Attention (HA) block [74] before feature-enhanced process, as shown in Fig. 5. This block refines boundary regions in $\mathbf{S}_{\text{pre}}$ by enlarging the foreground coverage area, and produces the blurred map $\mathbf{S}_{\text{opt}}$ of $\mathbf{S}_{\text{pre}}$. The process in HA can be formulated as follow:

$$\mathbf{S}_{\text{opt}} = MAX(P_{\min\_\max}(conv_g(\mathbf{S}_{\text{pre}}, k)), \mathbf{S}_{\text{pre}}), \tag{3}$$

where $conv_g(\cdot)$ is the Gaussian convolution operation to blur S_{pre} with size 32 and standard deviation k which is set to 4, $P_{\min\_\max}(\cdot)$ is a normalization operation, and $MAX(\cdot)$ is a maximum function. Subsequently, we conduct feature-enhanced process for the middle-level features, *i.e.* $\mathbf{f}'_{1,3}$ and $\mathbf{d}_3$, utilizing $\mathbf{S}_{\text{opt}}$ to suppress the distractors of $\mathbf{f}'_{1,3}$ and $\mathbf{d}_3$ via element-wise multiplication. Then, we achieve residual connection with original $\mathbf{f}'_{1,3}$ and $\mathbf{d}_3$ via element-wise summation. A channel weighting process (CW) for modulating RGB features is additionally adopted. In this way, more discriminative channel-feature maps will be strengthened adaptively. Considering depth maps are less informative, the channel-wise self-weighed operation is just adopted for RGB feature maps. These operations in TB module are formulated as follow:

$$\mathbf{f}_{\text{oe}} = \mathbf{f}'_{1,3} \otimes \mathbf{S}_{\text{opt}} \oplus \mathbf{f}'_{1,3}, \tag{4}$$

$$\mathbf{d}'_3 = \mathbf{d}_3 \otimes \mathbf{S}_{\text{opt}} \oplus \mathbf{d}_3, \tag{5}$$

$$\mathbf{f}_{2,3} = \mathbf{f}_{\text{oe}} \otimes G(\mathbf{f}_{\text{oe}}) \oplus \mathbf{f}_{\text{oe}}, \tag{6}$$

where $\mathbf{f}_{\text{oe}}$ and $G(\cdot)$ denote RGB feature maps enhanced by $\mathbf{S}_{\text{opt}}$ and a global pooling layer, respectively. Notably, there are no parameters to learn in TB module. Therefore, TB module does not cost additional computation consumption.

The features of TB module are shown in Fig. 6. Compared with the original middle-level features $\mathbf{f}'_{1,3}$ and $\mathbf{d}_3$, the refined features $\mathbf{f}_{2,3}$ and $\mathbf{d}'_3$ focus more on the salient regions and have less distractors. The salient objects in $\mathbf{f}_{2,3}$ are rendered more evenly and have fewer cavities compared to $\mathbf{f}_{oe}$. These visually prove that the combination of HA block and CW process is reasonable and effective. Therefore, TB module can purify features to boost the detection ability of refinement subnetwork to better locate salient objects and sharpen the preliminary results.

3.4. Implementation details

Total Training Loss: The overall framework is trained end-to-end with pixel-level ground truths, and the total training loss L_{Total} is defined as the sum of two subnetworks' losses:

$$L_{\text{Total}} = L_{ce}(\mathbf{S}_{\text{pre}}, \mathbf{GT}) + L_{ce}(\mathbf{S}_{\text{final}}, \mathbf{GT}), \tag{7}$$

where **GT** is ground truth and L_{ce} is the cross-entropy loss. It is worth noting that the size of two subnetworks' outputs are both 88×88, thus they need to be resized to 352×352 for loss evaluation.

Training Data Setting: Intending for a fair comparison with state-of-the-art models, following [28,29,37], we take 1400 image pairs and 650 image pairs from NJU2K [39] and NLPR [13], respectively, as training data. And limited by the scale of training data, the training data is augmented by mirror flipping and rotation operations. All image pairs are resized to 352×352 for training and testing.

Network Training: The experiments are implemented on Pytorch 1.2.0 framework [75] by adapting a NVIDIA GTX 2080TI GPU (11G memory). We utilize Adam [76] optimizer for training, and set the batch size to 10, the initial learning rate to 10^{-4}, the decay rate for learning rate to 0.1, and the number of epoch to 59. The parameters of the CNN blocks are initialized by VGG16 [73], and the parameters of other convolution layers are initialized through the default setting of the Pytorch. The parameters of RGB stream and depth stream are not shared, and the parameters of the prediction subnetwork and the refinement subnetwork are also not shared. The training loss converges after 55 epochs, taking about 10 h.

4. Experimental results

4.1. Datasets

To verify the effectiveness and robustness of the proposed CCNet, we carry out comparison experiments on six benchmark datasets, NJU2K [39], NLPR [13], STEREO [16], DES [38], SIP [37], and LFSD [77]. STEREO is the first proposed dataset in this field with 1000 image pairs totally, while the quality of depth map is poor. Its gray distributions cannot match the object shape of corre-

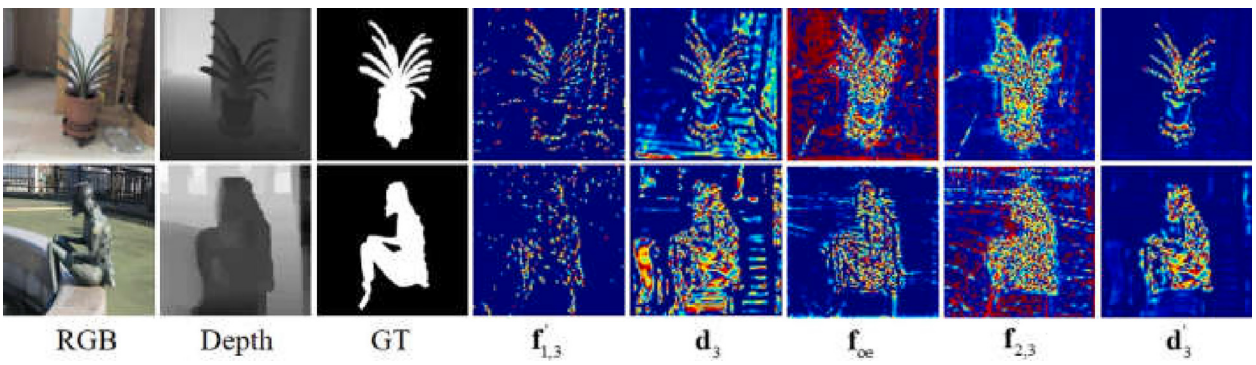


Fig. 6. Feature visualization in TB module. $\mathbf{f}'_{1,3}$ and $\mathbf{d}_3$ are the input RGB feature and depth feature of TB module, respectively. $\mathbf{f}_{oe}$ is produced by CW process. $\mathbf{f}_{2,3}$ and $\mathbf{d}'_3$ are the output RGB feature and depth feature of TB module, respectively.

sponding RGB image. LFSD and DES are small datasets, and the images of each have uniform resolutions. LFSD includes 60 pairs of indoor and 40 pairs of outdoor images. DES includes 135 pairs of indoor images. Compared with DES, the scenarios of LFSD are more complex and challenging. NJU2K and NLPR consist of 1985 and 1000 pairs of images, respectively. SIP dataset is built for salient person detection in the wild with 1000 pairs of images.

4.2. Evaluation metrics

To conduct a comprehensive performance evaluation, six evaluation metrics are employed in this paper, including Precision-Recall (P-R) curve, S-measure (S_α) [78], maximum F-measure (F_β) [79], weighted F-measure [80], maximum E-measure (E_ε) [81] and mean absolute error (MAE) [82]. P-R curve is plotted by 256 pairs of precision and recall which come from the comparison between GT and binary maps generated by thresholding pixels of a saliency map with a series of fixed integers from 0 to 255. S-measure is used to evaluate region-aware and object-aware structural similarity between a saliency map and a GT. E-measure captures matching information of image-level statistics and local pixel jointly. F-measure is defined to measure the overall performance of the saliency map:

$$F_\beta = \frac{(1+\beta^2)\cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \tag{8}$$

where $\beta^2 = 0.3$, as the same as [20].

The direct estimation between GT and saliency map is given by MAE:

$$MAE = \frac{1}{N}|SM - GT|, \tag{9}$$

where N represents a total number of pixels in an image, and SM denotes saliency map.

4.3. Ablation study

To demonstrate the impact of CFC module and TB module on our CCNet network, we conduct detailed experiments on STEREO [16], LFSD [77] and SIP [37] datasets. All the variants are trained with the same setting as the complete CCNet.

4.3.1. CFC module

We evaluate contributions of the CFC module, the prediction subnetwork, and the refinement subnetwork in this section. Additionally, we evaluate the rationality of suppressing the interference of depth features in CFC module, and compare the effect of different numbers and levels of CFC module.

The effectiveness of CFC module. We generate a variant by removing three CFC modules from CCNet, named as w/o CFC. The results of w/o CFC are shown in Table 1. Compared with the complete CCNet, the performance of w/o CFC drops dramatically on three evaluated datasets (*e.g.* S_α: 0.908 → 0.881, 0.876 → 0.801, 0.886 → 0.846). This demonstrates the CFC module is important for CCNet. Besides, since the CFC module contains the main interactions between RGB images and depth maps, w/o CFC disables the depth stream of the proposed network. Therefore, w/o CFC is also known as w/o Depth. The sharp drop in performance further proofs the validity of depth map for SOD.

The effectiveness of suppressing the interference of depth features. Compared with the existing cross-feature complementary modules [21,23,24,28], we innovatively propose the interference suppression operation for depth features in the RGB-to-depth direction of the CFC module. To verify the effectiveness of this operation, we remove the connection of RGB-to-depth from CFC modules, and constructed the variant w/o RGB → Depth. The results of w/o RGB → Depth obviously worse than that of the CCNet (*e.g.* S_α: 0.908 → 0.901, 0.876 → 0.855, 0.886 → 0.876; M: 0.037 → 0.045, 0.062 → 0.072, 0.048 → 0.059). As shown in Fig. 7, although the salient objects are correctly highlighted by

Table 1
Ablation studies for CFC module on STEREO, LFSD and SIP datasets. The best results are marked as **red**. The variants are described in detail in Section 4.3.1.

Model	STEREO		LFSD		SIP	
	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
CCNet (Ours)	**0.908**	**0.037**	**0.876**	**0.062**	**0.886**	**0.048**
w/o CFC (w/o Depth)	0.881	0.062	0.801	0.103	0.846	0.068
w/o RGB→Depth	0.901	0.045	0.863	0.072	0.876	0.059
w/o Refine-CFC (w/o CFC-H)	0.903	0.039	0.863	0.069	0.878	0.053
w/o Pre-CFC (w/o CFC-L&M)	0.902	0.041	0.848	0.084	0.873	0.057
w/o CFC-L	0.903	0.040	0.869	0.068	0.880	0.051
w/o CFC-M&H	0.893	0.049	0.838	0.088	0.863	0.066
w/o CFC-M	0.896	0.044	0.853	0.076	0.870	0.055
w/o CFC-L&H	0.902	0.042	0.860	0.071	0.877	0.056

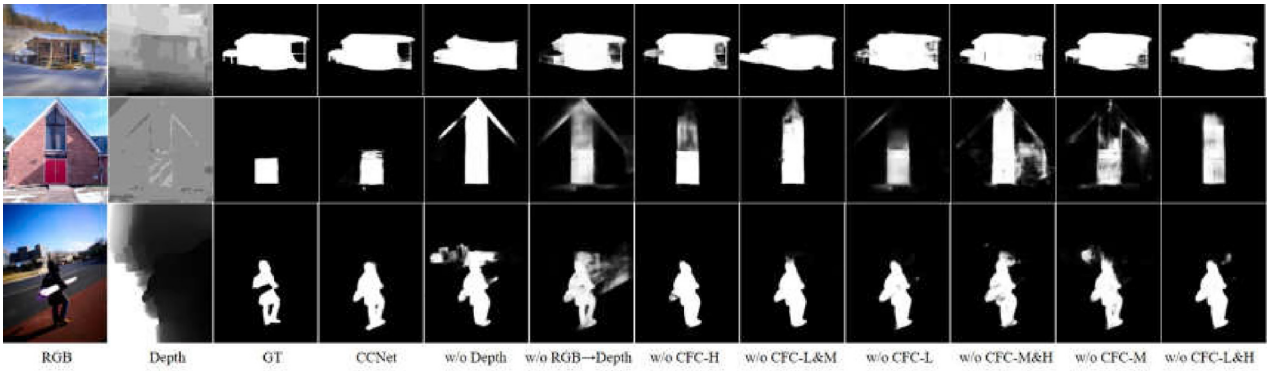


Fig. 7. Visual comparisons of CCNet with variants about CFC module.

Table 2
Ablation studies for TB module on STEREO, LFSD and SIP datasets. The best results are marked as **red**. The variants of TB module are described in Section 4.3.2.

Model	STEREO		LFSD		SIP	
	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
CCNet (Ours)	0.908	**0.037**	0.876	0.062	**0.886**	0.048
w/o TB-M	0.898	0.050	0.868	0.075	0.869	0.066
w/o CW	0.904	0.039	0.870	0.066	0.881	0.052
w/ 2TB (w/ 2Refine)	**0.911**	**0.037**	**0.880**	**0.061**	0.883	**0.046**
w/ TB-L	0.909	**0.037**	0.867	0.065	0.884	0.050
w/ TB-H	0.904	0.039	0.864	0.069	0.878	0.053

w/o RGB → Depth, the non-salient regions in saliency maps are contaminated by the poor-quality depth maps. This observation demonstrates the interference suppression operation indeed improves the performance to a certain extent.

The effectiveness of CFC module in the prediction and refinement subnetworks. CFC modules in the prediction and refinement subnetworks are removed in turn to construct the variants w/o Pre-CFC and w/o Refine-CFC, respectively. From Table 1, the performance of w/o Pre-CFC is significantly inferior than the complete CCNet on three datasets (*e.g.* S_α: 0.908 → 0.902, 0.876 → 0.848, 0.886 → 0.873). The performance of w/o Refine-CFC also decreases (*e.g.* S_α: 0.908 → 0.903, 0.876 → 0.863, 0.886 → 0.878). The fallen performance of both w/o Pre-CFC and w/o Refine-CFC confirms the necessity of CFC module in two subnetworks.

The different levels and the number of CFC modules. We take turns removing or keeping three levels of CFC module to construct six variants, including w/o CFC-L, w/o CFC-M&H, w/o CFC-M, w/o CFC-L&H, w/o CFC-H, and w/o CFC-L&M. As reported in Table 1, we observe that all six variants perform worse than CCNet, which indicates that the CFC module of each level is indispensable. Additionally, in terms of the performance degradation of each variant, the performance degradation of w/o CFC-M is more serious than w/o CFC-L and w/o CFC-H, and the degradation of w/o CFC-L&H is less than w/o CFC-M&H and w/o CFC-L&M. These all prove that CFC module plays a greater role in middle-level features than it does in low- and high-level features.

4.3.2. TB Module

The effectiveness of TB module. We discard TB module from CCNet and name the variant w/o TB-M, due that the interface of TB module is the middle-level feature. Removing TB module indicates the middle-level features are directly imported into refinement module without any enhanced operations. As reported in Table 2, the performance degradation of w/o TB-M (*e.g.* S_α: 0.908 → 0.898, 0.876 → 0.868, 0.886 → 0.869; M: 0.037 → 0.050, 0.062 → 0.075, 0.048 → 0.066) proves that the TB module can effectively boost the detection accuracy. As shown in Fig. 8, the example saliency maps clearly show that the CCNet captures salient boundaries quite well due to the effectiveness of TB module.

The effectiveness of channel weighting operation in the TB module. The variant w/o CW denotes the TB module without channel weighting operation. Compared with the complete CCNet, the performance of w/o CW slightly decreases (*e.g.* S_α: 0.908 → 0.904, 0.876 → 0.870, 0.886 → 0.881; M : 0.037 → 0.039, 0.062 → 0.066, 0.048 → 0.052). Although the channel weighting operation is simple and occupies less computation burden, it is an indispensable operation in the TB module.

The Number of TB Modules. The w/ 2 TB (w/ 2Refine) is a variant of integrating another one TB module behind the original refinement subnetwork. Due that the function of TB module is to bridge two subnetworks, we additionally add a refinement subnetwork behind the second TB module. With the computation cost increasing, the performance improvement of w/ 2 TB (w/ 2Refine) is subtle, as reported in Table 2. Therefore, considering both performance improvement and computation cost, one TB module is proper for our CCNet.

Selection of Optimization Feature. The w/ TB-L and w/ TB-H are constructed by replacing the optimization features of TB module with low- and high-level features, respectively. From Table 2, w/ TB-L performs differently on three datasets. On the whole, the performance slightly degrades (*e.g.* S_α: 0.908 → 0.909, 0.876 → 0.867, 0.886 → 0.884; M : 0.037 → 0.037, 0.062 → 0.061, 0.048 → 0.046). The performance of w/ TB-H descends on all tested datasets (*e.g.* S_α : 0.908 → 0.904, 0.876 → 0.864, 0.886 → 0.878; M : 0.037 → 0.039, 0.062 → 0.069, 0.048 → 0.053) due that the decoder in the refinement subnetwork integrates middle-level features with less discrimination.

4.4. Comparison with State-of-the-arts

To evaluate the performance of our CCNet for RGB-D SOD, we compare it with 13 state-of-the-art models, including CDCP [34], MDSF [31], DF [32], CTMF [22], PCF [21], AFNet [18], MMCI [23], TANet [12], CPFP [29], DMRA [20], D3Net [37], AGSD [44] and ICNet [28]. CDCP and MDSF are traditional models, and the rest of these models are based on CNNs.

4.4.1. Quantitative comparisons

We summarize S-measure [78], maximum F-measure [79], maximum E-measure [81] and MAE [82] of two predictions of our CCNet and other state-of-the-art models on NJU2K [39], LFSD [77], DES [38], NLPR [13], SIP [37] and STEREO [16] datasets in Table 3. Besides, we post PR curves and the weighted F-measure [80] of 8 advanced models, including AFNet [18], MMCI [23], TANet [12], CPFP [29], DMRA [20], D3Net [37], AGSD [44] and ICNet [28], in Fig. 9.

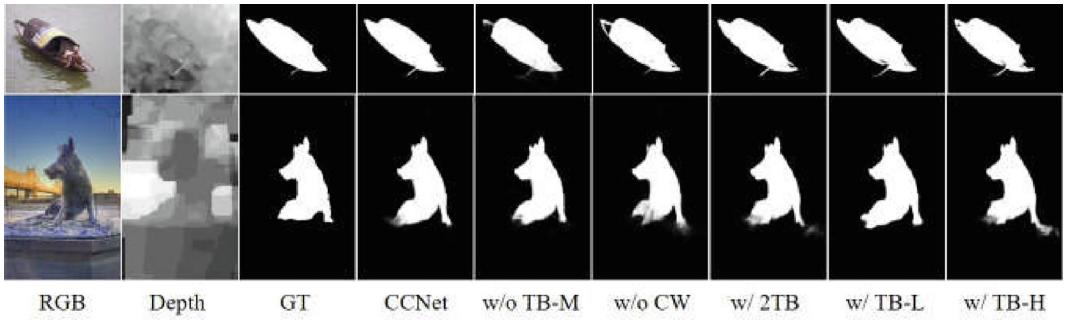
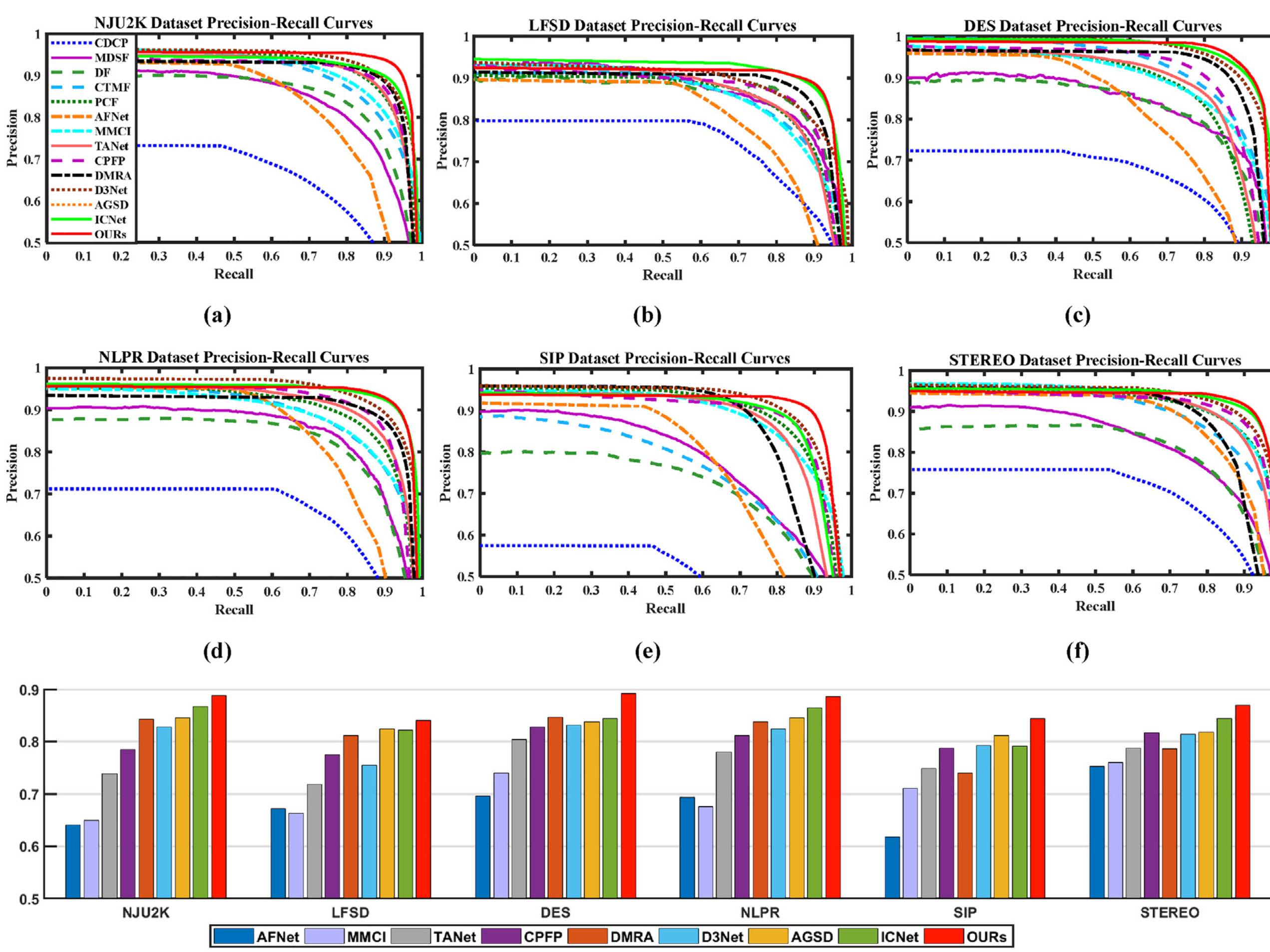


Fig. 8. Visual comparisons of CCNet with variants about TB module.

Table 3
Quantitative comparisons on 6 public datasets. ↑ indicates larger is better, and ↓ denotes smaller is better. The best result is marked **red**, the second place is marked **green**, and the third place is marked **blue**.

Model	NJU2K-T [39]				LFSD [77]				DES [38]				NLPR-T [13]				SIP [37]				STEREO [16]			
	S_α↑	F_β↑	E_ε↑	M↓	S_α↑	F_β↑	E_ε↑	M↓	S_α↑	F_β↑	E_ε↑	M↓	S_α↑	F_β↑	E_ε↑	M↓	S_α↑	F_β↑	E_ε↑	M↓	S_α↑	F_β↑	E_ε↑	M↓
CDCP17 [34]	.669	.621	.741	.180	.717	.703	.786	.167	.709	.631	.811	.115	.727	.645	.820	.112	.595	.505	.721	.224	.713	.664	.786	.149
MDSF17 [31]	.748	.775	.838	.157	.700	.783	.826	.190	.741	.746	.851	.122	.805	.793	.885	.095	.717	.698	.798	.167	.728	.719	.809	.176
DF17 [32]	.763	.804	.864	.141	.791	.817	.865	.138	.752	.766	.870	.093	.802	.778	.880	.085	.653	.657	.759	.185	.757	.757	.847	.141
CTMF18 [22]	.849	.845	.913	.085	.796	.791	.865	.119	.863	.844	.932	.055	.860	.825	.929	.056	.716	.694	.829	.139	.848	.831	.912	.086
PCF18 [21]	.877	.872	.924	.059.	.794	.779	.835	.112	.842	.804	.893	.049	.874	.841	.925	.044	.842	.838.	.901	.071	.875	.860	.925	.064
AFNet19 [18]	.772	.775	.853	.100	.738	.744	.815	.133	.770	.728	.881	.068	.799	.771	.879	.058	.720	.712	.819	.118	.825	.823	.887	.075
MMCI19 [23]	.858	.852	.915	.079	.787	.771	.839	.132	.848	.822	.928	.065	.856	.815	.913	.059	.833	.818	.897	.086	.873	.863	.927	.068
TANet19 [12]	.878	.874	.925	.060	.801	.796	.847	.111	.858	.827	.910	.046	.886	.863	.941	.041	.835	.830	.895	.075	.871	.861	.923	.060
CPFP19 [29]	.878	.877	.923	.053	.828	.826	.872	.088	.872	.846	.923	.038	.888	.867	.932	.036	.850	.851	.903	.064	.879	.874	.925	.051
DMRA19 [20]	.886	.886	.927	.051	.847	.856	.900	.075	.900	.888	.943	.030	.899	.879	.947	.031	.806	.821	.875	.085	.835	.847	.911	.066
D3Net19 [37]	.895	.889	.932	.051	.832	.819	.864	.099	.904	.885	.943	.030	.906	.885	.946	.034	.864	.862	.903	.063	.891	.881	.930	.054
AGSD20 [44]	.892	.890	.927	.055	.851	.851	.886	.084	.902	.880	.938	.034	.915	.899	.950	.030	.872	.880	.916	.063	.889	.881	.928	.055
ICNet20 [28]	.894	.891	.926	.052	.868	.871	.903	.071	.920	.913	.960	.027	.923	.908	.952	.028	.854	.857	.890	.069	.903	.898	.942	.045
OURs-pre	.901	.909	.927	.045	.864	.885	.872	.083	.906	.915	.933	.027	.902	.907	.933	.027	.861	.891	.882	.068	.886	.897	.893	.061
OURs	.917	.929	.948	.037	.875	.894	.907	.061	.922	.936	.966	.022	.926	.922	.966	.023	.886	.902	.925	.048	.908	.913	.944	.037



(g) Weighted F-measure comparison with 8 advanced models on 6 datasets

Fig. 9. Quantitative comparisons of the proposed CCNet with 8 state-of-the-art models. (a)-(f) are P-R curves.

As can be seen from Table 3, the proposed CCNet shows superior performance, especially on the NJU2K and SIP datasets. In addition, the preliminary prediction of CCNet (*i.e.* OURs-pre) presents competitive performance. The S-measure scores of the proposed CCNet beyond the second place D3Net (except for OURs-pre) by 2.2% on NJU2K. Even on the STEREO dataset which has lots of poor-quality depth maps, CCNet is still in the lead. This indicates that the integrity of the detected salient objects by the proposed CCNet is better than that of other models. As shown in Fig. 9, on each dataset, the convex points of our CCNet are above

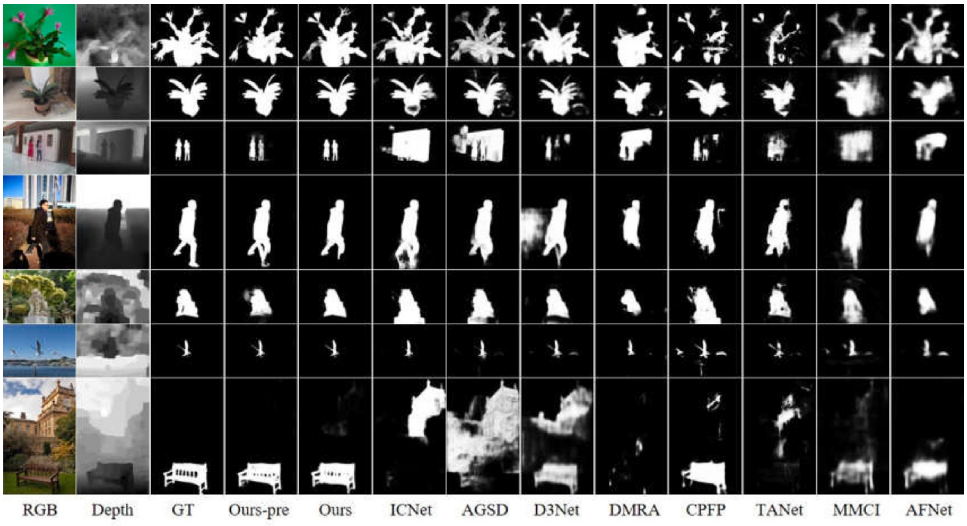


Fig. 10. Visual comparisons of the proposed CCNet with eight state-of-the-art models.

other state-of-the-art models, and the weighted F-measure scores of our CCNet are still higher than other models.

4.4.2. Visual comparisons

Some saliency map examples of eight advanced models and the proposed CCNet are presented in Fig. 10 for visual comparison. The saliency predictions of the proposed CCNet are superior than other models and show three advantages: the detected salient regions have sharper contrasts with non-salient regions; the structures of detected objects have finer edges than other results; the generated saliency maps have less distractors, even confronting challenging scenarios, such as low contrast, complex scene, background disturbance, and confused depth map. For example, in the first row of Fig. 10, the flowers in the RGB image and palms in depth map are complemented to detect the complete cactus. In the fifth row, the trees of depth map are suppressed by the RGB image and the effective information of stone lion in the depth map is extracted to assist RGB features in turn. These are attributed to the circular complementary manner in the CFC modules. In the sixth row, the tiny edge of seagull can still be detected through the TB module even the saliency map is up-sampled from 88×88. Compared with our preliminary results, the salient objects in our final saliency maps are highlighted evenly with sharper object boundaries, and some falsely highlighted regions in the preliminary results are also suppressed (*e.g.* the 3rd, 4th and 5th rows). These improvements all benefit from the effectiveness of the TB module and the rational utilization of the predict-refine scheme in our CCNet. Visual comparisons are consistent with the quantitative results in Table 3, which further demonstrates the superiority of our CCNet.

5. Conclusion

In this paper, we propose a predict-refine architecture based CCNet for accurate RGB-D SOD. CCNet consists of the prediction subnetwork and the refinement subnetwork, and it is equipped with the novel CFC modules and a TB module. Compared with the existing cross-modal feature complementary module, the CFC module is embedded with additional RGB-to-depth integration to enhance depth features, and further boosts the mutual complementary of RGB and depth features. With the TB module, the prediction subnetwork and the refinement subnetwork are bridged to improve the integrity of CCNet and more discriminated features are promoted to reach more accurate SOD. Ablation results of multiple variants verify the effectiveness and rationality of CFC module and TB module. Comparison results demonstrate the proposed CCNet outperforms 13 state-of-the-art models on six benchmark datasets, and it is more robust to RGB image with poor-quality depth map.

CRediT authorship contribution statement

Zhen Bai: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft. **Zhi Liu:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition, Resources. **Gongyang Li:** Validation, Writing - review & editing. **Linwei Ye:** Formal analysis, Data curation. **Yang Wang:** Writing - review & editing.

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.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 61771301.

References

[1] J. Han, D. Zhang, G. Cheng, N. Liu, D. Xu, Advanced deep-learning techniques for salient and category-specific object detection: a survey, IEEE Signal Process, Magazine 35 (1) (2018) 84–100.
[2] Y. Yang, J. Sun, H. Li, Z. Xu, ADMM-CSNet: A deep learning approach for image compressive sensing, IEEE Trans. Pattern Anal. Mach. Intel. 42 (3) (2019) 521–538.
[3] A. Serrano, I. Kim, Z. Chen, S. Diverdi, D. Gutierrez, A. Hertzmann, B. v, Motion parallax for 360° RGBD video, IEEE Trans. Visual. Comp. Graph. 25 (5) (2019) 1817–1827.
[4] L. Ye, M. Rochan, Z. Liu, Y. Wang, Cross-modal self-attention network for referring image segmentation, in: in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 10502–10511.
[5] G. Li, Z. Liu, R. Shi, W. Wei, Constrained fixation point based segmentation via deep neural network, Neurocomputing 368 (2019) 180–187.

[6] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, Y. Wang, Salient object segmentation via effective integration of saliency and objectness, IEEE Trans. Multimedia 19 (8) (2017) 1742–1756.
[7] G. Li, Z. Liu, R. Shi, Z. Hu, W. Wei, Y. Wu, M. Huang, H. Ling, Personal fixations-based object segmentation with object localization and boundary preservation, IEEE Trans. Image Process. 30 (2021) 1461–1475.
[8] M. Huang, Z. Liu, L. Ye, X. Zhou, Y. Wang, Saliency detection via multi-level integration and multi-scale fusion neural networks, Neurocomputing 364 (2019) 310–321.
[9] Y. Wu, Z. Liu, X. Zhou, Saliency detection using adversarial learning networks, J. Vis. Commun. Image Represent. 67 (2020) 102761.
[10] C. Lang, T.V. Nguyen, H. Katti, K. Yadati, M.S. Kankanhalli, S. Yan, Depth matters: Influence of depth cues on visual saliency, in: Proceedings of European Conference on Computer Vision (ECCV), Springer, 2012, pp. 101–115.
[11] A. Ciptadi, T. Hermans, J.M. Rehg, An in depth view of saliency, in: Proceedings of British Machine Vision Conference (BMVC), Springer, 2013, pp. 9–13.
[12] H. Chen, Y. Li, Three-stream attention-aware network for RGB-D salient object detection, IEEE Trans. Image Process. 28 (6) (2019) 2825–2835.
[13] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: a benchmark and algorithms, in: Proceedings of European Conference on Computer Vision (ECCV), Springer, 2014, pp. 92–109.
[14] Y. Piao, X Li, M. Zhang, J. Yu, H. Lu, Saliency detection via depth-induced cellular automata on light field, IEEE Trans. Image Process. 29 (2020) 1879–1889.
[15] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, H. Ling, Hierarchical alternate interaction network for RGB-D salient object detection, IEEE Trans. Image Process. 30 (2021) 3528–3542.
[16] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 454–461.
[17] X. Fan, Z. Liu, G. Sun, Salient region detection for stereoscopic images, in: Proceedings of International Conference on Digital Signal Processing (ICDSP), ACM, 2014, pp. 454–458.
[18] N. Wang, X. Gong, Adaptive fusion for RGB-D salient object detection, IEEE Access 7 (2019) 55277–55284.
[19] D. Feng, N. Barnes, S. You, C. Mccarthy, Local background enclosure for RGB-D salient object detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 2343–2350.
[20] Y. Piao, W. Ji, J. Li, M. Zhang, H. Lu, Depth-induced multi-scale recurrent attention network for saliency detection, in: Proceedings of the International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 7254–7263.
[21] H. Chen, Y.F. Li, Progressively complementarity-aware fusion network for RGB-D salient object detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 3051–3060.
[22] 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. Sys. Man. Cyber. 48 (11) (2018) 3171–3183.
[23] 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 Recognition 86 (2019) 376–385.
[24] H. Chen, Y. Li, D. Su, Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection, in: Proceedings of Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 6821–6826.
[25] Z. Liu, W. Zhang, P. Zhao, A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection, Neurocomputing 387 (2020) 210–220.
[26] 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.
[27] Y. Ding, Z. Liu, M. Huang, R. Shi, X. Wang, Depth-aware saliency detection using convolutional neural networks, J. Vis. Commun. Image Represent. 61 (2019) 1–9.
[28] G. Li, Z. Liu, H. Ling, ICNet: Information conversion network for RGB-D based salient object detection, IEEE Trans. Image Process. 29 (2020) 4873–4884.
[29] J. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, Li X., L. Zhang, Contrast prior and fluid pyramid integration for RGBD salient object detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 3927–3936.
[30] C. Zhu, X. Cai, K. Huang, T.-H. Li, G. Li, PDNet: Prior-model guided depth-enhanced network for salient object detection, in: Proceedings of International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 199–204.
[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] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks, IEEE Trans. Image Process. 27 (1) (2018) 38–49.
[34] G. Li, W. Wang, R. Wang, An innovative salient object detection using center-dark channel prior, in: Proceedings of International Conference on Multimedia and Expo (ICME), IEEE, 2017, pp. 1509–1515.
[35] X. Lu, W. Wang, J. Shen, Y.-W. Tai, S.C.H. Hoi, Learning video object segmentation from unlabeled videos, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 8960–8970.
[36] X. Lu, W. Wang, C. Ma, J. Shen, F. Porikli, See more, know more: Unsupervised video object segmentation with co-attention siamese networks, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 3623–3632.
[37] D.-P. Fan, Z. Lin, J. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, M.-M. Cheng, Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks, arXiv preprint arXiv: 1907.06781v1 (2019).
[38] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, in: Proceedings of International Conference on Internet Multimedia Computing and Service (ICIMCS), ACM, 2014, pp. 23–27.
[39] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in: Proceedings of International Conference on Image Processing (ICIP), IEEE, 2014, pp. 1115–1119.
[40] J. Ren, X. Gong, L. Yu, W. Zhou, M.Y. Yang, Exploiting global priors for RGB-D saliency detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 25–32.
[41] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, L. Qing, Stereoscopic saliency model using contrast and depth-guided-background prior, Neurocomputing 275 (2018) 2227–2238.
[42] A. Wang, M. Wang, RGB-D salient object detection via minimum barrier distance transform and saliency fusion, IEEE Signal Process. Let. 24 (5) (2017) 663–667.
[43] W. Wang, J. Shen, Y. Yu, K.-L. Ma, Stereoscopic thumbnail creation via efficient stereo saliency detection, IEEE Trans. on Visual. and Com. Graph. 23 (8) (2017) 2014–2027.
[44] G. Li, X. Zhou, C. Gong, Z. Liu, J. Zhang, Attention-guided RGBD saliency detection using appearance information, Image Vis. Comput. 95 (2020) 103888.
[45] L. Itti, C.K.E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern. Anal. Mach. Intel. 20 (11) (1998) 1254–1259.
[46] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: Proceedings of International Conference on Computer Vision (ICCV), IEEE, 2009, pp. 2106–2113.
[47] A.C. Ngo, L.-M. Ang, G. Qiu, K.P. Seng, Multi-scale visual attention & saliency modelling with decision theory, in: Proceedings of International Conference on Image Processing (ICIP), IEEE, 2013, pp. 216–220.
[48] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Proceedings of Neural Information Processing Systems (NIPS), NIPS Foundation, 2007, pp. 545–552.
[49] A. Borji, L. Itti, State-of-the-art in visual attention modeling, IEEE Trans. Pattern Anal. Mach. Intel. 35 (1) (2013) 185–207.
[50] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, A. Borji, Revisiting video saliency prediction in the deep learning era, IEEE Trans. Pattern Anal. Mach. Intel. 43 (1) (2021) 220–237.
[51] E. Vig, M. Dorr, D. Cox, Large-scale optimization of hierarchical features for saliency prediction in natural images, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 2798–2805.
[52] W. Wang, J. Shen, Deep visual attention prediction, IEEE Trans. Image Process. 27 (5) (2018) 2368–2378.
[53] Z. Wang, Z. Liu, W. Wei, H. Duan, SalED: Saliency prediction with a pithy encoder-decoder architecture sensing local and global information, Image and Vision Computing, http://dx.oi.org/10.1016/j.imavis.2021.104149.
[54] A. Borji, Saliency prediction in the deep learning era: Successes and limitations, IEEE Trans. Pattern Anal. Mach. Intel. 43 (2) (2021) 679–700.
[55] K. Shi, K. Wang, J. Lu, L. Lin, PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 2115–2122.
[56] X. Hou, J. Harel, C. Koch, Image signature: Highlighting sparse salient regions, IEEE Trans Pattern Anal. Mach. Intel. 34 (1) (2012) 194–201.
[57] Z. Liu, X. Zhang, S. Luo, O.Le Meur, Superpixel-based spatiotemporal saliency detection, IEEE Trans. Circuits Syst. Video Technol. 24 (9) (2014) 1522–1540.
[58] Z. Liu, J. Li, L. Ye, G. Sun, L. Shen, Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation, IEEE Trans. Circuits Syst. Video Technol. 27 (12) (2017) 2527–2542.
[59] J. Shen, Y. Du, W. Wang, X. Li, Lazy random walks for superpixel segmentation, IEEE Trans. Image Process. 23 (4) (2014) 1451–1462.
[60] J. Shen, J. Peng, X. Dong, L. Shao, F. Porikli, Higher order energies for image segmentation, IEEE Trans. Image Process. 26 (10) (2017) 4911–4922.
[61] J. Shen, X. Dong, J. Peng, X. Jin, L. Shao, F. Porikli, Submodular function optimization for motion clustering and image segmentation, IEEE Trans. Neural Netw Learn Syst. 30 (9) (2019) 2637–2649.
[62] W. Wang, J. Shen, L. Shao, F. Porikli, Correspondence driven saliency transfer, IEEE Trans. Image Process. 25 (11) (2016) 5025–5034.
[63] J. Ren, Z. Liu, X. Zhou, G. Sun, C. Bai, Saliency integration driven by similar images, J. Vis. Commun. Image Represent. (2018) 227–236.
[64] A. Borji, M.-M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Trans. Image Process. 24 (12) (2015) 5706–5722.
[65] G. Li, Y. Yu, Visual saliency based on multiscale deep features, IEEE Trans. Image Process. 25 (11) (2016) 5012–5024.
[66] W. Wang, J. Shen, X. Dong, A. Boji, R. Yang, Inferring salient objects from human fixations, IEEE Trans. Pattern Anal. Mach. Intel. 42 (8) (2020) 1913–1927.
[67] W. Wang, J. Shen, M.M. Cheng, L. Shao, An iterative and cooperative top-down and bottom-up inference network for salient object detection, in: Proceedings

of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 5961–5970.
[68] J. Zhao, J. Liu, D. Fan, Y. Cao, J. Yang, M.-M. Cheng, EGNet: Edge guidance network for salient object detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 8779–8788.
[69] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, BASNet: Boundary-aware salient object detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 7479–7489.
[70] W. Wang, S. Zhao, J. Shen, S.C.H. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 1448–1457.
[71] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, R. Yang, Salient object detection in the deep learning era: An in-depth survey, IEEE Trans. Pattern Anal. Mach. Intel. (2021), https://doi.org/10.1109/TPAMI.2021.3051099.
[72] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, P. Heng, R3Net: Recurrent residual refinement network for saliency detection, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann, 2018, pp. 684–690.
[73] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Proceedings of International Conference on Learning Representations (ICLR), Elsevier, 2015.
[74] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 3907–3916.
[75] K.M. Chen, E.M. Cofer, J. Zhou, O.G. Troyanskaya, Selene: A pytorch-based deep learning library for sequence-level data, bioRxiv preprint bioRxiv: (2018) 438291.
[76] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: Proceedings of International Conference on Learning Representations (ICLR), Elsevier, 2015, pp. 1–15.
[77] N. Li, J. Ye, Y. Ji, H. Ling, J. Yu, Saliency detection on light field, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 2806–2813.
[78] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: a new way to evaluate foreground maps, in: Proceedings of the International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 4558–4567.
[79] R. Achanta, S.S. Hemami, F.J. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 1597–1604.
[80] R. Margolin, L. Zelnikmanor, A. Tal, How to evaluate foreground maps, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 248–255.
[81] D.-P. Fan, C. Gong, Cao Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann, 2018, pp. 698–704.
[82] F. Perazzi, P. Krahenbuhl, Y. Pritch, A. Hornung, Saliency filters: contrast based filtering for salient region detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 733–740.

Zhen Bai received the B.E. degree from Wuhan Huaxia University of Technology, Wuhan, China, in 2016, the M. S. degree from Zhengzhou University of Light Industry, Zhengzhou, China, in 2019, and is currently pursuing the Ph.D. degree with School of Communication and Information Engineering, Shanghai University, Shanghai, China. Her research interests include machine learning and saliency detection.

Zhi Liu received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China, in 1999, 2002, and 2005, respectively. He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. From Aug. 2012 to Aug. 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 200 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision and multimedia communication. He was a TPC member/session chair in ICIP 2017, PCM 2016, VCIP 2016, ICME 2014, WIAMIS 2013, etc. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an area editor of *Signal Processing: Image Communication* and served as a guest editor for the special issue on *Recent Advances in Saliency Models, Applications and Evaluations* in *Signal Processing: Image Communication*. He is a senior member of IEEE.

Gongyang Li received the B.E. degree from Shanghai Normal University, Shanghai, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai. His research interests include image/video object segmentation and saliency detection.

Linwei Ye received the B.E. degree from Hangzhou Dianzi University, Hangzhou, China, in 2013, the M.E. degree from Shanghai University, Shanghai, China, in 2016, and is currently working toward the Ph.D. degree in computer science at the University of Manitoba, Winnipeg, MB, Canada. His research interests include saliency model, salient object segmentation, and semantic segmentation.

Yang Wang received the B.Sc. degree from the Harbin Institute of Technology, Harbin, China, the M.Sc. degree from the University of Alberta, Edmonton, AB, Canada, and the Ph.D. degree from Simon Fraser University, Burnaby, BC, Canada, all in computer science. He was previously a NSERC Postdoc Fellow with the University of Illinois at Urbana-Champaign, Champaign, IL, USA. He is currently an Associate Professor of computer science with the University of Manitoba, Winnipeg, MB, Canada. His research interests include computer vision and machine learning.