

FCMNet: Frequency-aware cross-modality attention networks for RGB-D salient object detection [☆]

Xiao Jin ^a, Chunle Guo ^b, Zhen He ^a, Jing Xu ^{a,*}, Yongwei Wang ^c, Yuting Su ^d

^a College of Artificial Intelligence, Nankai University, Tianjin 300350, China

^b College of Computer Science, Nankai University, Tianjin 300350, China

^c Joint NTU-WeBank Research Centre on Fintech, Nanyang Technological University, 639798, Singapore

^d School of Electronic and Information Engineering, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:

Received 27 July 2021

Revised 26 January 2022

Accepted 3 April 2022

Available online 5 April 2022

Communicated by Zidong Wang

Keywords:

RGB-D salient object detection

Feature fusion

Attention mechanism

2D discrete cosine transform

ABSTRACT

RGB-D saliency detection aims to comprehensively use RGB images and depth maps to detect object saliency. This field still faces two challenges: 1) how to extract representative multimodal features and 2) how to effectively fuse them. Most of the previous methods in this field equally treat RGB and depth information as two modalities, while not considering the difference in the frequency domain of the two modalities, and may lose some complementary information. In this paper, we introduce the frequency channel attention mechanism into the fusion process. First, we design a frequency-aware cross-modality attention (FACMA) module to interweave adequate channel features and select representative features. In the FACMA module, we also propose a spatial frequency channel attention (SFCA) module to introduce more complementary information in different channels. Second, we develop a weighted cross-modality fusion (WCMF) module to adaptively fuse multimodality features by learning the content-dependent weight maps. Comprehensive experiments on several benchmark datasets demonstrate that the proposed framework outperforms seventeen state-of-the-art methods.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Salient object detection (SOD) aims to extract the most visually distinctive regions from the background in an image [1,2]. SOD is a widely used preprocessing tool that is beneficial for many visual applications, such as object detection [3], image quality assessment [4], image retrieval [5], image compression [6], image caption [7], and video tracking [8].

When the texture and color of the background are similar to those of the foreground, these algorithms cannot achieve satisfactory results. Therefore, researchers integrate depth information as auxiliary inputs to improve the performance of salient object detection. Since depth sensors, such as Kinect, have become affordable and easily available, a large number of RGB-depth (RGB-D) image pairs can be obtained. Thus, the RGB-D salient object detec-

tion (RGB-D SOD) problem has attracted increasing attention [9,10].

Motivation. As shown in Fig. 1, some existing methods can hardly obtain acceptable results in some challenging scenarios. A careful review of the existing approaches for RGB-D SOD shows that the following issues require further study and improvement [11]:

- (1) How to extract representative features from different modalities. Although attention mechanisms are employed for choosing the most representative features, most previous RGB-D SOD schemes only focus on one type of attention component [12]. The ability to select effective features in the attention mechanism has not been fully explored. In addition, most existing methods depend on spatial and channel attention. How to select suitable features from other perspective of the attention mechanism is still an open problem that has not been well studied in this visual task.
- (2) How to preserve complementary features from different modalities. Global average pooling (GAP) is a standard operation in several kinds of attention modules. However, its ability to capture complementary information is not ideal [13]. Unique characteristics in channel features may be lost

[☆] This work was supported in part by the Science and Technology Planning Project of Tianjin, China (17JCZDJC30700, 18ZXZNGX00310), the Tianjin Natural Science Foundation (19JCQNJC00300), and the Fundamental Research Funds for the Central Universities of Nankai University (632011192, 632111116).

* Corresponding author.

E-mail address: xujing@nankai.edu.cn (J. Xu).

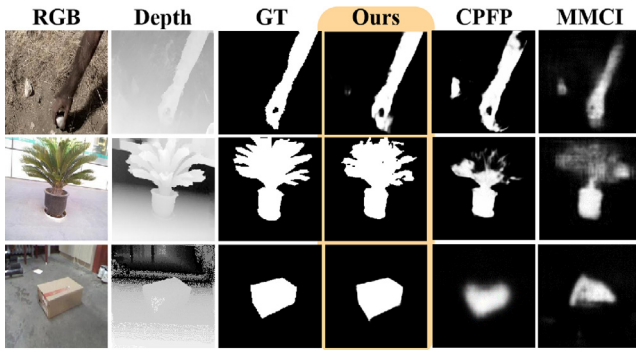


Fig. 1. Some results of the proposed model and some state-of-the-art models. These examples cover several challenging circumstances, including multiple objects, complex objects, and poor-quality depth maps.

in the process of calculating the mean value in the GAP operations. Thus, traditional attention mechanisms may not attain satisfactory experimental results in RGB-D SOD.

- (3) How to fuse these heterogeneous features. Fusion strategies by concatenation or element-wise summation operation do not consider the difference between two modalities. The results of these fusion operations may even be worse than those obtained using RGB information only, since some low-quality depth maps may negatively affect the RGB features [9]. Furthermore, these straightforward fusion operations do not take into account the content dependency of the multimodality data and the nonlinear representation ability of the network.

Contribution. Our core observation is that some previous cross-modality fusion methods may not preserve all of the complementary information in channel features [13]. Since deep networks are redundant, some channels may obtain the same information using GAP. The multispectral framework can extract more information from redundant channels because various frequency components explore different information. The proposed framework consists of three key modules: the frequency-aware cross-modality attention (FACMA) module, the spatial frequency channel attention (SFCA) module, and the weighted cross-modality fusion (WCMF) module. The main contributions of this article are as follows:

- A frequency-aware cross-modality attention network (FCMNet) is proposed, which is an end-to-end architecture designed for RGB-D SOD. Unlike previous methods that only consider spatial and channel attention, the proposed method explores this task from the perspective of the frequency domain. A novel network module called frequency-aware cross-modality attention (FACMA) is presented, which can extract discriminative features while maintaining the complementary components.
- We also develop a weighted cross-modality fusion (WCMF) module to adaptively incorporate multimodality features by weighing their importance. Different from previous models that lack cross-modal interactions and content dependency, our module considers those factors, and weakens the effects of low-quality depth information. Furthermore, using the nonlinear feature enhancement (NFE) unit, we enhance the nonlinear representation ability during the fusion process.
- We compare the proposed methods with 17 state-of-the-art approaches on eight widely used datasets. Without any preprocessing or postprocessing techniques, our method achieves the best performance under four evaluation metrics. In addition, extensive ablation studies are conducted, which demonstrate the effectiveness of the proposed modules.

The rest of the paper is organized as follows. [Section 2](#) reviews the related work on attention mechanisms and RGB-D SOD. [Section 3](#) presents the proposed network and several novel network modules designed in this work. In [Section 4](#), in addition to elaborating the datasets, evaluation metrics, and compared methods, we present and analyze the experimental results. Finally, the conclusion summarizes this paper in [Section 5](#).

2. Related Work

In this section, we review some prior art closely related to our work. These works can be divided into attention mechanisms and salient object detection. We also point out several existing problems in previous work that are addressed in this paper.

2.1. Attention Mechanism

The attention mechanism aims to focus on the features of interests and neglect interference, which has been widely applied in computer vision tasks. The attention mechanism can be roughly divided into three types: spatial attention, channel attention, and the combination of spatial and channel attention.

Hu et al. [14] designed a squeeze-and-excitation (SE) block to flexibly update channel-wise features by directly modeling the relationship between the channels. Inspired by the SE block, Roy et al. [15] introduced a concurrent spatial and channel squeeze & excitation (scSE) module, which combines the channel and spatial attention units in parallel. Woo et al. [16] presented a lightweight and general module called the convolutional block attention module (CBAM), which connects channel and spatial attention in series. Park et al. [17] proposed a bottleneck attention module (BAM) to indicate significant regions in a 3D attention map. Gao et al. [18] constructed a global second-order pooling (GSoP) block. Given an input tensor, the GSoP module first calculates the covariance matrix, and then performs convolution and activation operations to obtain the outputs.

Attention mechanisms are widely adopted in previous SOD methods [19–24]. To generate attentive global features, Liu et al. [20] proposed a multiscale global attention model for feature fusion. In the work [19], an attention mechanism is introduced to dynamically select the message propagation in graph neural networks (GNNs).

Although attention mechanisms have been employed for RGB-D SOD, some previous methods only adopted a single type of attention module [12]. The effect of the combination of attention modules has not been well explored. Furthermore, in the global pooling operation, which is widely used in the attention module, it is often difficult to retain complementary information of different modalities [13]. Thus, traditional attention components can hardly obtain satisfactory results in cross-modal tasks. To solve these problems, we design the FACMA module, which will be elaborated in [Section 3.1](#).

2.2. RGB-D Saliency Detection

Conventional RGB-D SOD methods usually regard the depth map as another channel in addition to RGB. Niu et al. [25] first introduced an additional depth cue for saliency detection from stereoscopic images. Peng et al. [26] proposed a multicontextual contrast approach for RGB-D SOD, and built the first large-scale RGB-D dataset for salient object detection. Feng et al. [27] noticed that high-contrast regions in the background may cause false positives, and they proposed local background enclosure (LBE) features to solve this problem. Song et al. [28] presented an RGB-D SOD approach based on multiscale discriminative saliency fusion

(MDSF) and bootstrap learning. Cong et al. [29] proposed a depth-guided transformation model (DTM), which consists of multilevel RGBD saliency initialization, depth-guided saliency refinement, and saliency optimization with depth constraints. However, the abovementioned methods depend heavily on handcrafted features and heuristic fusion, causing unsatisfactory results and a lack of generalizability in challenging scenarios.

In recent years, deep learning has been widely applied in RGB-D SOD [30–32]. Chen et al. [33] noticed the previous ambiguousness of fusing cross-modal data and proposed a multiscale multipath cross-modal interaction (MMCI) architecture for RGB-D SOD. Zhao et al. [34] noted that backbone models pretrained on ImageNet cannot extract optimal features from the depth channel, so they combine contrast prior with fluid pyramid networks. Piao et al. [35] presented an RGB-D SOD framework by residual connections and recurrent attention module. Fan et al. [9] found that low-quality depth maps affect the final results of RGB-D SOD, and consequently designed a deep depth-depurator network (D3Net) to automatically discard low-quality depth maps in the test phase. Li et al. [36] proposed the cmMS block for fusing RGB image and depth information adaptively. Zhang et al. [37] considered global location and local detail complementarities from RGB and depth modalities. In the work [38], a depth distiller (A2dele) is described by minimizing the distances between feature maps of the depth and RGB information. Liu et al. [22] proposed a novel unified model for both RGB and RGB-D SOD based on a pure transformer. For the first time, RGB-D SOD is solved from a new sequence-to-sequence perspective. Zhang et al. [39] developed probabilistic RGB-D saliency detection network based on conditional variational autoencoders, which models human annotation uncertainty and produces multiple saliency maps for each input. Liu et al. [24] utilized the self-attention to enhance longrange contextual dependencies, and further designed a selection attention and a residual fusion module to improve the performance.

Though the representation ability of deep learning is powerful [40], the satisfactory fusion scheme of two modalities is still a difficult task. Depth maps contain more contours, while RGB images convey rich detailed information, such as texture and color. The development of an effective method for fusing the two modality features with different statistical characteristics is vital for improving the performance of RGB-D SOD. Therefore, we introduce the WCMF module to solve this problem. The details of this module are presented in Section 3.2.

3. Methodology

The overall architecture of the FCMNet is a two-stream encoder-decoder neural network, as shown in Fig. 2. The input of the network is an RGB image \mathcal{I} and a depth image \mathcal{D} . We copy the depth image into three channels to make it have the same dimension as the RGB image, $\{\mathcal{I}, \mathcal{D}\} \in \mathbb{R}^{H \times W \times C}$, where H , W , and C represent height, width and the number of channels, respectively. The encoder consists of two symmetrical VGG-16 networks. The output features of the RGB branch and depth branch in each stage are defined as \mathcal{F}_{RGB}^i and \mathcal{F}_D^i , where $i \in \{1, 2, 3, 4, 5\}$. These feature maps are fed into a frequency-aware cross-modality attention (FACMA) module to obtain the corresponding enhanced features \mathcal{F}_{RGBE}^i and \mathcal{F}_{DE}^i . The enhanced features are fed to a weighted cross-modality fusion (WCMF) module to adaptively fuse the multimodal features. Then, the fused features are input to the subsequent cascaded decoder, which integrates multiscale features progressively. The decoder is composed of four atrous spatial pyramid pooling (ASPP) modules [41] to enrich the multiscale information. These ASPP blocks are augmented with a dense connection, which promotes the integration of depth and RGB features at dif-

ferent scales. During training, the feature maps are supervised by the ground truth and the edge ground truth. Note that we choose the second stage in VGG-16 to embed the edge supervision information, which is consistent with the previous work of RGB SOD [42].

3.1. Frequency-Aware Cross-Modality Attention

As is known, RGB images and depth maps contain various information. Depth maps contain more contours, while RGB images convey rich detailed information, such as texture and color. Conventional attention mechanisms based on GAP cannot maintain all frequency components from different modalities [13]. To preserve these complementary features, we analyze this problem from the perspective of the frequency domain. We design an FACMA module to automatically extract and strengthen complementary information in different modalities. In the FACMA module, two spatial frequency channel attention (SFCA) submodules are utilized to capture complementary information from the spatial and frequency domains. The feature maps of the RGB branch and depth branch pass through two symmetrical SFCA modules and then undergo elementwise multiplication to interweave different modality information. This process can be formulated as:

$$\mathcal{F}_{RGBE}^i = \text{SFCA}(\mathcal{F}_{RGB}^i) \otimes \mathcal{F}_D^i, \quad (1)$$

$$\mathcal{F}_{DE}^i = \text{SFCA}(\mathcal{F}_D^i) \otimes \mathcal{F}_{RGB}^i, \quad (2)$$

where \otimes represents elementwise multiplication and $\text{SFCA}(\cdot)$ is the proposed spatial frequency channel attention module.

As shown in Fig. 3, the proposed FACMA module contains two submodules called the SFCA module. Inspired by [15], we design a spatial frequency channel attention (SFCA) module to capture the complementary information from different modalities. As illustrated in Fig. 4, the SFCA module can be divided into two components. The first is the spatial attention module, which is used to accurately extract position information,

$$f_1 = (\sigma(\text{Conv}_{1 \times 1}(f_{in}))) \otimes f_{in}, \quad (3)$$

where $\sigma(\cdot)$ is the sigmoid function, $\text{Conv}_{1 \times 1}(\cdot)$ represents the 1×1 convolution operation, and \otimes refers to elementwise multiplication.

The other is the frequency channel attention (FCA) module [13], which is used to capture the response component of different channels to the salient area. The process of the FCA module can be described as:

$$f_2 = \text{FCA}(f_{in}) = \sigma(\text{FC}(\text{ReLU}(\text{FC}(\text{DCT}(f_{in})))))) \otimes f_{in}, \quad (4)$$

where $\sigma(\cdot)$ is the sigmoid function, $\text{FC}(\cdot)$ denotes the fully connected layer, $\text{ReLU}(\cdot)$ is the rectified linear unit (ReLU) activation, and $\text{DCT}(\cdot)$ refers to the 2D discrete cosine transform (DCT), which outputs a matrix with the same size of the input. The 2D DCT is mathematically defined as follows:

$$\text{DCT}(f_{in}) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} f_{in} \cos\left(\frac{\pi h}{H}\left(x + \frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(y + \frac{1}{2}\right)\right), \quad (5)$$

$$h \in \{0, 1, \dots, H-1\}, w \in \{0, 1, \dots, W-1\},$$

where H and W are the height and width of the input feature map, respectively. In [13], the authors proved that FCA module can preserve richer features than the common channel attention module.¹

Then, we use the element addition operation to combine the two components. This process can be described as:

$$f_{out} = f_1 \oplus f_2, \quad (6)$$

¹ The codes and implementation details for the FCA layer are available at: <https://github.com/cfzd/FcaNet>.

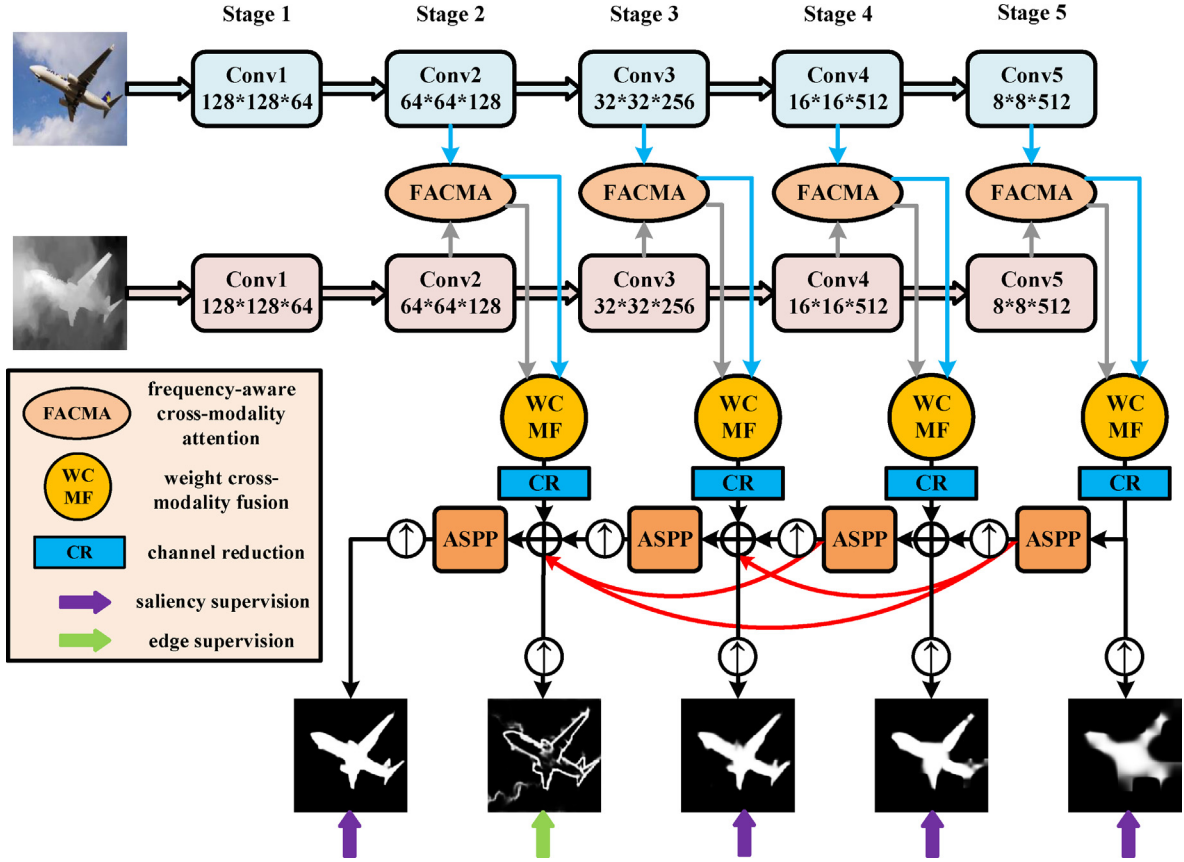


Fig. 2. The overall architecture of the FCMNet is a two-stream encoder-decoder neural network. The encoder consists of two symmetrical VGG-16 networks. The feature maps of the RGB branch and depth branch in each stage are fed into a frequency-aware cross-modality attention (FACMA) module to obtain the corresponding enhanced features. The enhanced features are fed to a weighted cross-modality fusion (WCMF) module to adaptively fuse the multimodal features. The decoder is composed of four atrous spatial pyramid pooling (ASPP) modules. These ASPP blocks are augmented with a dense connection, which is represented by red lines. “ \oplus ” and “ \uparrow ” represent the elementwise addition and upsampling, respectively. During training, the feature maps are supervised by the ground truth and the edge ground truth. The ground truths and the input images have the same resolution.

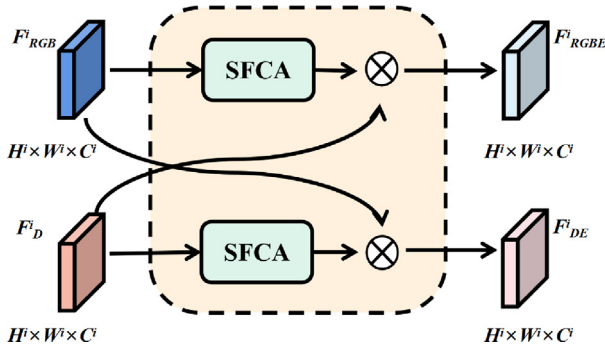


Fig. 3. The proposed frequency-aware cross-modality attention (FACMA) module. In this figure, “ \otimes ” represents the elementwise multiplication operation. In the i -th stage, the dimension of feature maps inputted to FACMA is denoted as $H^i \times W^i \times C^i$.

where \oplus denotes elementwise addition and f_{out} is the final output of the SFCA module.

3.2. Weighted Cross-Modality Fusion

The previous multimodal fusion strategies in RGB-D SOD often adopt elementwise summation or concatenation. Redundant and misleading features are easily involved, thereby reducing the complementarity between RGB images and depth maps. Meanwhile, these methods ignore the image content information, which is of great significance for salient object detection [43–45]. Moreover,

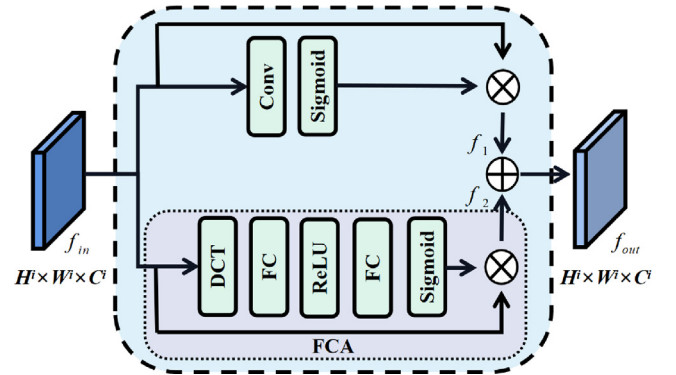


Fig. 4. The proposed spatial frequency channel attention (SFCA) module. In this figure, “ \otimes ” and “ \oplus ” represent the elementwise multiplication and addition, respectively. In the i -th stage, the dimension of feature maps inputted to SFCA is denoted as $H^i \times W^i \times C^i$.

earlier work neglects the nonlinear representation ability of the neural network during the fusion process. Thus, we propose the weighted cross-modality fusion (WCMF) module to solve the above problems.

The structure of the WCMF module is shown in Fig. 5. For brevity, we first denote a nonlinear feature enhancement (NFE) unit before introducing the WCMF module. The NFE unit is composed of a convolutional layer, a batch normalization (BN) layer, and a

rectified linear unit (ReLU) activation function. This unit can be described as:

$$\text{NFE}_{1 \times 1}(\cdot) = \text{ReLU}(\text{BN}(\text{Conv}_{1 \times 1}(\cdot))), \quad (7)$$

$$\text{NFE}_{3 \times 3}(\cdot) = \text{ReLU}(\text{BN}(\text{Conv}_{3 \times 3}(\cdot))), \quad (8)$$

where $\text{Conv}_{1 \times 1}(\cdot)$ and $\text{Conv}_{3 \times 3}(\cdot)$ refer to the 1×1 and 3×3 convolution, respectively. The WCMF module first uses the NFE operation to enhance the two-stream input features and concatenates them:

$$\bar{f}_1 = \text{NFE}_{1 \times 1}(\mathcal{F}_{RGBE}^i), \quad (9)$$

$$\bar{f}_2 = \text{NFE}_{1 \times 1}(\mathcal{F}_{DE}^i), \quad (10)$$

$$\bar{f}_3 = \text{Cat}[\bar{f}_1; \bar{f}_2], \quad (11)$$

where $\text{Cat}[\cdot; \cdot]$ denotes the concatenation operation. Then, we can obtain two concatenated weight maps by the fused feature map \bar{f}_3 :

$$[W_R; W_D] = \text{NFE}_{3 \times 3}(\text{NFE}_{3 \times 3}(\bar{f}_3)), \quad (12)$$

where W_R and W_D are the weight maps corresponding to RGB and depth feature, respectively. Finally, the output of the WCMF module is computed as:

$$\bar{f}_{out} = W_R \otimes \bar{f}_1 \oplus W_D \otimes \bar{f}_2 \oplus W_R \otimes \bar{f}_1 \otimes W_D \otimes \bar{f}_2. \quad (13)$$

As suggested in [46], equipped with nonlinear representation ability enhancement and adaptively content-dependent weight maps, the proposed fusion module can capture complementary information from different modalities. The NFE unit can strengthen the representation properties of the neural networks. The weight maps can assign different attention to feature maps in different modalities and locations. Thus, the proposed fusion module can effectively and adaptively fuse the representative features from different modalities.

3.3. Decoder

As we can see in Fig. 2, the feature maps outputted from WCMF modules have various resolutions and channel numbers. Therefore, we compress the hierarchical features into the same channel number. After receiving features from the WCMF modules, the channel numbers of these feature maps are all converted to 64. This channel reduction process is conducted by convolution and ReLU activation, which is denoted as the “CR” modules in Fig. 2. There are two advantages of this attempt. First, a small number of feature channels is friendly to memory usage and computational consumption. Second, the same number of channels facilitates the elementwise operations.

By channel reduction and resolution upsampling, the output channel numbers and resolutions in each ASPP module are completely identical with the input ones. In Fig. 2, the red lines denote

dense connections. In each dense connection, the feature maps are upsampled to keep the same resolutions.

3.4. Loss Function

To better preserve the salient edge features, we added additional edge supervision information in the second stage of the decoder. It is widely accepted that low-level features tend to learn boundary information. In the first stage of VGG-16, the convolution layer is too close to the input and the receptive field is too small [42]. Therefore, we extract edge features from the second stage of the backbones. Since the positive and negative samples of edge information are imbalanced, we use the cross-entropy loss with weights [42], which is defined as:

$$\mathcal{L}^{(2)}(E, G_E, W_{\mathcal{E}}) = -W_{\mathcal{E}} \cdot \sum_j [G_{Ej} \log(E_j) + (1 - G_{Ej}) \log(1 - E_j)], \quad (14)$$

where j indicates the pixel index, E is the predicted feature map of the edge, and G_E is the ground truth of the edge. The weight is defined as follows:

$$W_{\mathcal{E}} = \mu \cdot E_+ + v \cdot E_-, \quad (15)$$

$$\mu = \frac{\sum_j E_-}{\sum_j E_- + \sum_j E_+}, \quad (16)$$

$$v = \frac{1.1 \cdot \sum_j E_+}{\sum_j E_- + \sum_j E_+}, \quad (17)$$

where E_+ and E_- represent the salient edge and background pixel set, respectively. In other stages, traditional cross-entropy loss is used to calculate the loss between the predicted saliency map and the ground truth, and is defined as:

$$\mathcal{L}^{(i)}(S, G) = -\sum_j [G \log(S_j) + (1 - G) \log(1 - S_j)], \quad i \in \{1, 3, 4, 5\}, \quad (18)$$

where S is the predicted saliency map and G is the ground truth, i indicates the index of stage in VGG-16 networks. Finally, the total loss can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}^{(2)}(E, G_E, W_{\mathcal{E}}) + \sum_i \mathcal{L}^{(i)}(S, G), \quad i \in \{1, 3, 4, 5\}. \quad (19)$$

All the ground truths G and G_E are of the same resolution as the input images. To keep the resolution consistent, we upsample the feature maps to the same size of the ground truths before comput-

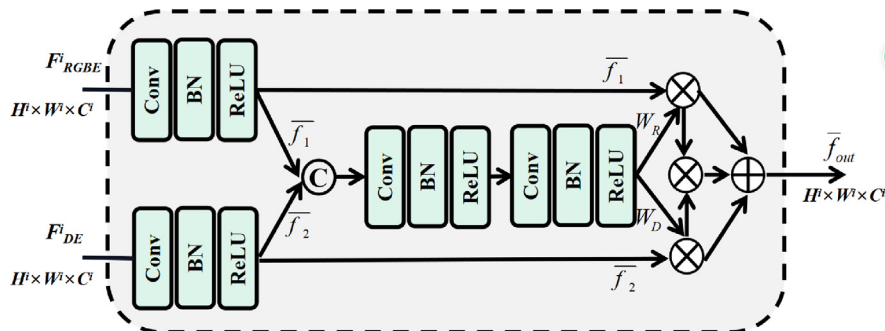


Fig. 5. The proposed weighted cross-modality fusion (WCMF) module. In this figure, “C” represents the concatenation operation. “ \otimes ” and “ \oplus ” represent the elementwise multiplication and addition, respectively. In the i -th stage, the dimension of feature maps inputted to WCMF is denoted as $H^i \times W^i \times C^i$.

ing the loss in each stage. This process is conducted by bilinear interpolations.

3.5. Discussion

(1) Discussion on the novelty of FACMA. Although attention mechanisms have been widely used in RGB-D SOD, the proposed FACMA module is inherently different from them in both motivation and structure. First, some previous literatures only consider a single type of attention modules [12]. However, our model combines the advantages of spatial and frequency channel attention. Second, most of existing attention components are rely on the GAP operation, which is incapable to extract the complex information for various inputs [13]. For example, RGB images include more high-frequency components, e.g., details, textures; while depth maps contain more low-frequency components, e.g., flat areas. If we only use GAP for cross-modality fusion, it is equivalent to the lowest frequency components of DCT [13], ignoring many other potentially useful frequency components. Thus, it will loss some complementary information in RGB-D SOD. To alleviate this loss, we reasonably embed the FCA layer into the current attention mechanism. This attempt can preserve the unique characteristics in channel features. As a result, the complementary information from different modalities can be maintained.

(2) Discussion on the novelty of WCMF. In some previous methods of RGB-D SOD, fusion modules are implemented by concatenation or elementwise summation operation. The results of these fusion operations are often unsatisfactory, since the RGB features may be contaminated by some low-quality depth maps [9]. To design a fusion strategy, we should consider the content dependence between the two modalities. As mentioned in Section 3.2, by calculating the weight maps W_R and W_D in Equ. 12, the WCMF module can extract content correlated knowledge to guide the fusion process. Furthermore, unlike some related works [46], we not only study the content dependence between the two modalities, but also consider the expression ability of deep learning. It is well known in the computer vision community that the nonlinear representation ability of neural networks can be enhanced by the combination of a convolutional layer, a BN layer, and a ReLU layer. In this paper, we insert these layers, which are denoted as the NFE units for brevity, into the proposed fusion module. Based on the above two considerations, our WCMF module can acquire better results.

4. Experiments and Results

4.1. Set up

Datasets. We evaluate our method on eight benchmark datasets. **NLPR** [26] comprises 1,000 images from 11 types of indoor and outdoor scenes. Of these, 650 images are used as the training set, and the remaining images are used as the test set. **RGBD135** [47] contains 135 image pairs from seven indoor scenes with only one object in each image. This dataset is also called the DES dataset in some reports in the literature. **STEREO** [25] consists of 1,000 pairs of binocular images with coarse depth quality, and the depth images are generated by an optical method. **NJUD** [48] includes 1,985 image pairs collected from indoor/outdoor environments and stereo movies. The depth maps are calculated from the stereo images. In this dataset, 1,400 samples are used as the training set, and the remaining samples are used as the test set. **SSD** [49] is a small-scale dataset with 400 samples. These images are high-resolution with the size of 960×1080 . The quality of the corresponding depth maps is relatively poor in this dataset. **LFSD** [50] is composed of 100 all-focus RGB images, the corresponding depth

maps captured by the Lytro light field camera, and the pixelwise ground truth masks. This dataset is designed for light field saliency detection. **SIP** [9] includes 929 images captured by Huawei Mate10 with high quality depth maps and annotations. The resolution of samples is fixed to 992×744 . **ReDWeb-S** [51] involves 3,179 images, which cover 332 scenes and 432 objects. The images in this dataset usually have high depth quality and large object size.

Following the generally accepted setting in [9], we trained our model on a combined subset. It consists of 1,400 samples from the NJU2K [48] dataset and 650 samples from the NLPR [26] dataset.

Evaluation Metrics. For quantitative comparison, we adopt four widely-used metrics: S-measure ($S_\alpha, \alpha = 0.5$) [52], max F-measure ($F_\beta, \beta^2 = 0.3$) [53], mean absolute error (MAE) [54] and max E-measure (E_ξ) [55]. In the following definition, j denotes the pixel coordinate.

MAE [54] measures the l_1 distance between the predicted saliency map S and the ground truth G and is defined as:

$$\text{MAE}(S, G) = \frac{1}{K} \sum_{j=1}^K |S_j - G_j|, \quad (20)$$

where K is the total number of pixels.

The predictions are binarized with multiple thresholds for the F-measure and E-measure. The F-measure F_β [53] computes the harmonic mean of precision and recall, which is calculated as:

$$F_\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \quad (21)$$

where β^2 is set to 0.3 as suggested in previous work. This setting will enhance the effect of precision. Considering that each binary threshold will calculate a F_β score, we report the maximum F_β score across all thresholds.

Since the MAE and F-measure neglect the structure information, we also employ a structure measure S_α [52]. It provides an evaluation method for continuous saliency prediction without binarization. The S-measure S_α combines the object-aware (S_{obj}) and region-aware (S_{reg}) structural similarity and is computed as follows:

$$S_\alpha = \alpha \cdot S_{obj} + (1 - \alpha) \cdot S_{reg}, \quad (22)$$

where $\alpha \in [0, 1]$ is the balance parameter and we set $\alpha = 0.5$ as the default [52].

The E-measure E_ξ [55] is a perceptual evaluation based on the human cognitive system. It can assess the local pixel level and global image level similarity between the predicted saliency map and the ground truth. The specific definition is as follows:

$$E_\xi = \frac{1}{K} \sum_{j=1}^K \xi(M, G), \quad (23)$$

where M is the binary mask converted from a saliency map and ξ is the enhanced alignment matrix [55].

Implementation Details. We conduct our experiments using the PyTorch toolbox on an NVIDIA 2080Ti GPU. The well-known VGG-16 [64] pretrained on the ImageNet database [65] is utilized as the backbone of both RGB and depth streams. The input depth maps are repeated three times in the channel direction to obtain the standard input format of VGG-16. All the input images are resized to 256×256 pixels. To prevent overfitting, these input images are augmented by random flipping, clipping, and rotation.

The whole proposed network is trained end-to-end by using the stochastic gradient descent (SGD) optimizer [66] with the batch size of 4 for 73 epochs. The model needs approximately 600,000 training iterations for convergence, which takes nearly 20 h. The learning rate is fixed to $1e-10$ in the entire training procedure.

The momentum is set to 0.99, and the weight decay is assigned as 0.0005. During the whole detection framework, our method does not adopt any preprocessing or postprocessing strategies to

improve performance. When testing, the FPS of our method is around 55.

Table 1

Quantitative results compared with seventeen RGB-D SOD methods. “-” means that the results are unavailable since the authors did not release them. “*” means that the methods are trained with the NJUD, NLPR, and DUT-RGBD training sets. \uparrow (\downarrow) indicates the larger (smaller), the better. The best and the second best results are highlighted in bold and underline, respectively.

Model	NLPR [26]				RGBD135 [47]				NJU2K [48]			
	$S_x \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow	$S_x \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow	$S_x \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow
FCMNet(ours)	0.916	0.908	0.949	0.024	0.905	0.913	0.949	0.025	0.901	0.907	0.929	<u>0.044</u>
ASIF(21TCYB)[56]	0.884	<u>0.900</u>	0.822	0.030	0.535	0.473	0.624	0.109	0.889	<u>0.900</u>	0.921	0.047
DQSF(21NC)[57]	0.900	0.884	0.938	0.034	0.879	0.863	0.931	0.036	0.892	0.891	0.928	0.051
MCMFNet(21SP)[11]	0.905	0.885	0.938	0.040	0.903	0.877	0.934	0.036	0.889	0.882	0.923	0.061
*A2dele(20CVPR)[38]	0.881	0.881	0.945	0.028	0.884	0.870	0.920	0.029	0.869	0.873	0.916	0.051
*SSF(20CVPR)[37]	<u>0.914</u>	0.896	0.935	<u>0.026</u>	<u>0.904</u>	0.884	0.941	<u>0.026</u>	<u>0.899</u>	0.896	0.931	0.043
D3Net(20TNNLS)[9]	0.905	0.885	0.945	0.033	<u>0.904</u>	0.885	<u>0.946</u>	0.030	0.886	0.886	0.927	0.051
*DMRA(19ICCV)[35]	0.899	0.879	<u>0.947</u>	0.031	0.900	<u>0.888</u>	0.943	0.030	0.893	0.887	<u>0.930</u>	0.051
CPFP(19CVPR)[34]	0.888	0.867	0.932	0.036	0.872	0.846	0.923	0.038	0.879	0.877	0.926	0.053
TANet(19TIP)[58]	0.886	0.863	0.941	0.041	0.858	0.828	0.910	0.046	0.878	0.874	0.925	0.060
MMCI(19PR)[33]	0.856	0.815	0.913	0.059	0.848	0.822	0.928	0.065	0.858	0.852	0.915	0.079
PCF(18CVPR)[59]	0.874	0.841	0.925	0.044	0.842	0.804	0.893	0.049	0.877	0.872	0.924	0.059
CTMF(17TCYB)[60]	0.860	0.825	0.929	0.056	0.863	0.844	0.932	0.055	0.849	0.845	0.913	0.085
DF(17TIP)[61]	0.802	0.782	0.782	0.080	0.744	0.761	0.867	0.094	0.764	0.800	0.865	0.137
MDSF(17TIP)[28]	0.805	0.793	0.885	0.095	0.741	0.746	0.856	0.090	0.748	0.775	0.838	0.157
CDCP(17ICCV)[62]	0.732	0.651	0.825	0.108	0.709	0.631	0.811	0.115	0.668	0.615	0.740	0.180
LBE(16CVPR)[27]	0.776	0.758	0.866	0.073	0.703	0.788	0.890	0.208	0.700	0.746	0.807	0.149
DCMC(16SPL)[63]	0.729	0.656	0.795	0.112	0.707	0.666	0.773	0.111	0.690	0.723	0.803	0.167
Model	STERE [25]				LFSD [50]				SSD [49]			
	$S_x \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow	$S_x \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow	$S_x \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow
FCMNet(ours)	0.899	0.904	0.939	0.043	0.862	0.883	0.903	<u>0.068</u>	<u>0.855</u>	0.860	<u>0.903</u>	0.055
ASIF(21TCYB)[56]	0.869	<u>0.894</u>	0.926	0.050	0.814	0.858	0.861	0.090	0.849	<u>0.846</u>	0.888	0.059
DQSF(21NC)[57]	<u>0.897</u>	0.888	0.932	0.048	0.844	0.839	0.884	0.086	-	-	-	-
MCMFNet(21SP)[11]	-	-	-	-	-	-	-	-	0.842	0.824	0.902	0.075
*A2dele(20CVPR)[38]	0.879	0.879	0.928	<u>0.044</u>	0.837	0.836	0.880	0.074	0.807	0.815	0.870	0.068
*SSF(20CVPR)[37]	0.893	0.889	0.936	<u>0.044</u>	<u>0.859</u>	<u>0.866</u>	<u>0.900</u>	0.066	0.844	0.845	0.899	<u>0.057</u>
D3Net(20TNNLS)[9]	0.886	0.886	<u>0.938</u>	0.047	0.826	0.810	0.861	0.073	0.857	0.834	0.910	0.059
*DMRA(19ICCV)[35]	0.889	0.878	0.929	0.054	0.823	0.841	0.886	0.087	0.857	0.821	0.892	0.058
CPFP(19CVPR)[34]	0.879	0.874	0.925	0.051	0.828	0.826	0.872	0.088	0.807	0.766	0.852	0.082
TANet(19TIP)[58]	0.871	0.861	0.923	0.060	0.801	0.796	0.847	0.111	0.840	0.810	0.897	0.063
MMCI(19PR)[33]	0.873	0.863	0.927	0.068	0.787	0.771	0.838	0.132	0.813	0.781	0.882	0.082
PCF(18CVPR)[59]	0.875	0.860	0.925	0.064	0.794	0.778	0.835	0.112	0.841	0.804	0.892	0.062
CTMF(17TCYB)[60]	0.848	0.831	0.912	0.086	0.795	0.791	0.864	0.119	0.773	0.721	0.851	0.098
DF(17TIP)[61]	0.751	0.757	0.847	0.142	0.784	0.814	0.864	0.141	0.743	0.734	0.828	0.143
MDSF(17TIP)[28]	0.728	0.719	0.846	0.176	0.694	0.791	0.819	0.197	0.673	0.703	0.779	0.192
CDCP(17ICCV)[62]	0.713	0.664	0.786	0.149	0.717	0.703	0.786	0.167	0.683	0.683	0.683	0.683
LBE(16CVPR)[27]	0.660	0.633	0.787	0.250	0.736	0.726	0.804	0.208	0.621	0.619	0.736	0.278
DCMC(16SPL)[63]	0.731	0.740	0.819	0.148	0.753	0.817	0.856	0.155	0.704	0.711	0.786	0.169
Model	SIP [9]				ReDWeb-S [51]							
	$S_x \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow	$S_x \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow				
FCMNet(ours)	0.858	0.881	<u>0.912</u>	<u>0.062</u>	0.679	0.675	<u>0.756</u>	0.157				
ASIF(21TCYB)[56]	0.373	0.250	0.552	0.269	0.435	0.392	0.567	0.291				
DQSF(21NC)[57]	-	-	-	-	-	-	-	-				
MCMFNet(21SP)[11]	-	-	-	-	-	-	-	-				
*A2dele(20CVPR)[38]	0.826	0.832	0.890	0.070	0.641	0.603	0.670	0.160				
*SSF(20CVPR)[37]	0.874	<u>0.880</u>	0.921	0.053	0.595	0.558	0.710	0.189				
D3Net(20TNNLS)[9]	0.806	0.821	0.875	0.085	0.689	<u>0.673</u>	0.768	<u>0.149</u>				
*DMRA(19ICCV)[35]	<u>0.864</u>	0.861	0.910	0.063	0.592	0.579	0.721	0.188				
CPFP(19CVPR)[34]	0.850	0.851	0.903	0.064	<u>0.685</u>	0.645	0.744	0.142				
TANet(19TIP)[58]	0.835	0.830	0.870	0.075	0.656	0.623	0.741	0.165				
MMCI(19PR)[33]	0.833	0.818	0.897	0.086	0.660	0.641	0.754	0.176				
PCF(18CVPR)[59]	0.742	0.838	0.901	0.071	0.655	0.627	0.743	0.166				
CTMF(17TCYB)[60]	0.716	0.694	0.829	0.139	0.641	0.607	0.739	0.204				
DF(17TIP)[61]	0.653	0.657	0.565	0.185	0.595	0.579	0.683	0.233				
MDSF(17TIP)[28]	0.717	0.698	0.645	0.167	-	-	-	-				
CDCP(17ICCV)[62]	0.595	0.505	0.683	0.224	-	-	-	-				
LBE(16CVPR)[27]	0.727	0.751	0.651	0.200	0.637	0.629	0.730	0.253				
DCMC(16SPL)[63]	0.683	0.618	0.598	0.186	0.427	0.348	0.549	0.313				

Table 2

Speeds and sizes of the proposed method and some other typical methods.

Method	Params (MB) ↓	Speed (FPS) ↑
MMCI (19PR) [33]	930	19
CPFP (19CVPR) [34]	278	6
DMRA (19ICCV) [35]	147	22
FCMNet (Ours)	196.65	55

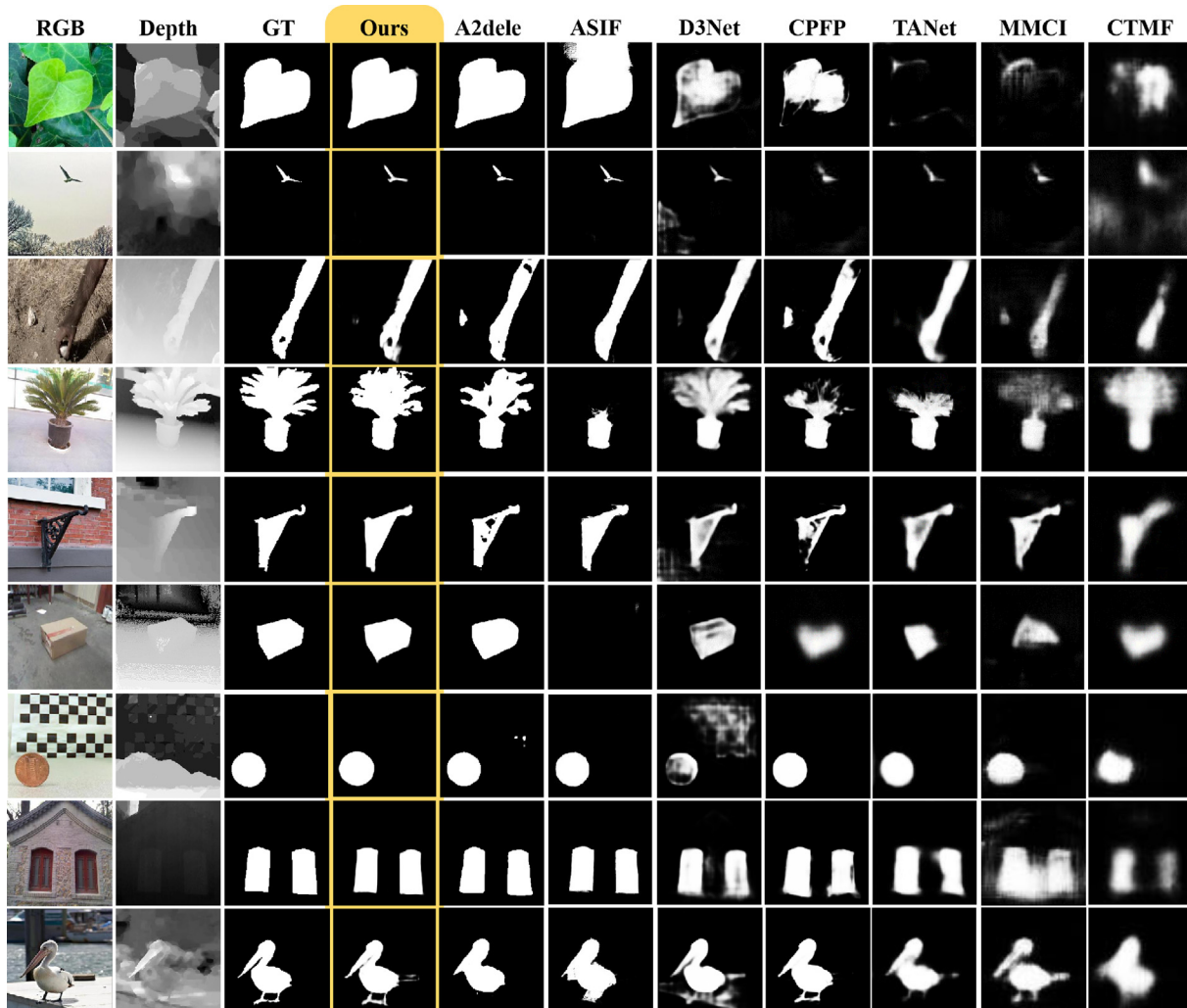
4.2. Comparison with SOTAs

In this subsection, we compare the proposed method with seventeen state-of-the-art RGB-D salient object detection methods, including four traditional methods: MDSF [28], CDCP [62], LBE [27], and DCMC [63], and thirteen DNN-based methods: MCMFNet [11], ASIF [56], DQSF [57], A2dele [38], SSF [37], D3Net [9], DMRA [35], CPFP [34], TANet [58], MMCI [33], PCF [59], CTMF [60], and DF [61]. For fair comparison, we report the quantitative results and saliency maps provided by the authors directly or generate these results using the corresponding codes with default parameters.

(1) Quantitative analysis: The quantitative comparison results under four different evaluation metrics on eight data sets are reported in Table 1. The comparison methods are presented from top to bottom according to the year of publication. For most eval-

uation metrics, e.g., F_β , S_α and E_ξ , a higher value indicates higher model effectiveness. On the contrary, the opposite is true for the MAE. Most comparison methods are trained with NJU2K and NLPR. In contrast, A2dele [38], SSF [37] and DMRA [35] are trained with the combination of NJU2K, NLPR and DUT. We mark these three methods with asterisks (“**”) to show the difference in their training datasets.

In this subsection, we analyze the quantitative results from two aspects, evaluation metrics and datasets. As shown in Table 1, the proposed framework leads to performance improvements according to several criteria. More concretely, our method consistently exceeds all the SOTAs in terms of F_β on all datasets. Compared with that of the second best method under the max F-measure, the performance gain of our method reaches 0.8% for NJU2K, 0.9% for NLPR, 2.8% for RGBD135, 1.7% for SSD, 1.1% for STERE, and 2.0% for LFSD. Based on this experimental observation, we can infer that the proposed FCMNet dependably detects salient objects rather than wrongly predicting background as salient regions. This finding provides a convincing demonstration that our modal is reliable. In addition, our method obtains the best S_α results on five datasets, e.g., NJU2K, NLPR, RGBD135, STERE, and LFSD, and the third best S_α results on one dataset, e.g., SSD. In terms of MAE, the proposed FCMNet outperforms other SOTA methods on four datasets, e.g., NLPR, RGBD135, SSD, and LFSD, and performs almost similarly to

**Fig. 6.** Qualitative comparison of the state-of-the-art RGB-D SOD methods and our approach.

the best method on two datasets, e.g., 0.044 (ours) v.s. 0.043 (SSF) in NJU2K, 0.068 (ours) v.s. 0.066 (SSF) in LFSD.

From the perspective of datasets, the RGBD135 dataset contains high-quality depth maps, while LFSD consists of some low-quality depth information. Our framework suppresses most methods on these two datasets, which indicates the generalization ability and robustness of FCMNet on datasets with different qualities of depth information. We note that the training set of A2dele [38], SSF [37] and DMRA [35] includes images from the DUT dataset [35]. DUT and LFSD are both collected with a Lytro Illum camera, which makes the samples in these two datasets similar. However, our results on LFSD still demonstrate competitive performance against the methods trained on DUT. We also achieve comparable performance in SIP dataset [9] and ReDWeb-S dataset [51]. The above analysis of the quantitative results proves the effectiveness of this method for improving performance in RGB-D SOD.

Besides the detection performance, we also compare the speed and parameter size of our FCMNet with some other RGB-D SOD methods. As listed in Table 2, our model contains relatively fewer parameters, which means that it is less prone to cause overfitting during training, and more convenient to run on portable devices. Although the parameters of DMRA [35] are fewer than our model, the test speed of the proposed FCMNet is faster than DMRA. As shown in the third column of Table 2, the proposed method is faster than other typical methods with a speed of 55 FPS.

(2) Qualitative analysis: To analyze our results more intuitively, some visualization results are shown in Fig. 6. These examples cover several challenging circumstances, including large objects, small objects, multi-objects, complex objects, and poor-quality depth maps. For instance, the first two rows show the capability of our model to detect large and small objects. Among all of the methods, only the proposed FCMNet can produce a complete structure and sharp boundaries. The image in the third row is chal-

lenging, because a stone is present in the background that has similar appearance to the foreground stone. Some methods will be disturbed by the stone in the background, e.g., CPFP [34], while our method can successfully separate the salient objects in the foreground. The last image is a typical case of low-quality depth maps. Our method achieves superior visual results compared with other approaches, e.g., D3Net [9]. Note that D3Net is designed for the case of low-quality depth maps in RGB-D SOD. These examples prove that our method is robust to noise in low-quality depth cues.

4.3. Ablation Study

In this subsection, we evaluate the effectiveness of the key components in the proposed neural networks. In detail, we analyze 1) the contribution of the FACMA, 2) the advantage of the SFCA, 3) the significance of the WCMF, and 4) the usefulness of edge supervision. In each comparative experiment, only one component is changed. Then, we retrain the neural networks with the same training protocol as before, and record the test results. Both qualitative and quantitative comparisons are discussed in this subsection. The visual examples in Fig. 7 are taken from the same eight datasets described in Section 4.1. Since the experimental phenomenon and the quantitative trends found on these eight datasets are similar, we only reported the results on NJU2K [48], NLP [26] and RGBD135 [47] in Table 3.

(1) The contribution of the FACMA module. In the proposed FCMNet, the FACMA module plays a significant role in performance enhancement. To evaluate the effectiveness of this module, we omit it in the whole neural network. The output feature maps from each stage in VGG-16 are directly connected to the WCMF module. The quantitative consequence of this change, denoted as “w/o FACMA”, is reported in Table 3. As shown in the table, different evaluation indicators on these three datasets show evidently lower

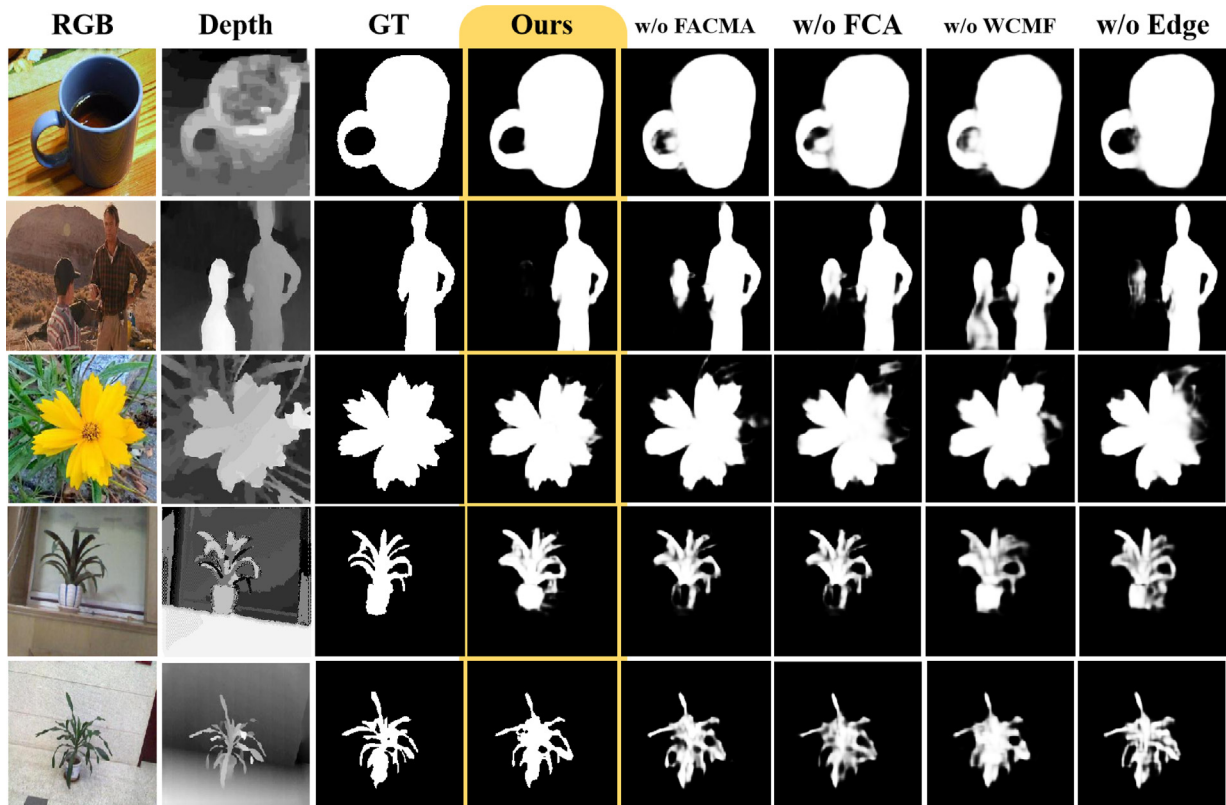


Fig. 7. The visual results of ablation analysis.

Table 3

Ablation analysis for the proposed FCMNet on the NLP, NJU2K and RGBD135 datasets. \uparrow (\downarrow) indicates the larger (smaller), the better. The best and the second best results are highlighted in bold and underline, respectively.

Models	NLP [26]				NJU2K [48]				RGBD135 [47]			
	$S_z \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow	$S_z \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow	$S_z \uparrow$	$F_\beta \uparrow$	$E_c \uparrow$	MAE \downarrow
w/o FACMA	0.904	0.898	0.934	0.031	0.889	0.896	0.914	0.048	0.893	0.903	0.937	0.031
w/o FCA	0.908	0.902	0.936	0.031	0.894	0.901	0.923	0.049	0.899	0.904	0.938	0.030
w/o WCMF (add)	0.898	0.888	0.924	0.035	0.884	0.889	0.906	0.052	0.894	0.900	0.946	0.032
w/o WCMF (concat)	0.891	0.883	0.929	0.035	0.886	0.896	0.919	0.050	0.853	0.859	0.912	0.043
w/o Edge	0.908	0.901	0.936	0.030	0.897	0.906	0.924	0.047	0.897	0.898	0.944	0.031
FCMNet	0.916	0.908	0.947	0.026	0.899	0.907	0.927	0.044	0.905	0.909	0.949	0.026

values. Some visual examples are presented in the fifth column of Fig. 7. The network structure without the FACMA module is susceptible to the interferences caused by multiple objects, e.g., the child in the second image. The above ablation analysis proves the contribution of the FACMA module.

(2) The advantage of the SFCA module. In this subsection, we conduct experiments to confirm the advantage of SFCA. By replacing the FCA unit with the common channel attention module, the proposed SFCA is converted to an scSE module [15]. In Table 3, without a frequency-aware attention mechanism, the proposed model cannot effectively preserve the complementary features from different modalities. As shown in the sixth column of Fig. 7, the results of “w/o FCA” bring some backgrounds into the final salient results, because the complementary information between the two modalities is not used effectively.

(3) The significance of the WCMF module. The WCMF module aims to enhance and fuse the features from different modalities. In this subsection, we replace the WCMF with an elementwise addition operation or a concatenation operation to verify the effect of this component. This attempt, denoted as “w/o WCMF (add)” and “w/o WCMF (concat)”, does not achieve better performance in Table 3. Some examples are illustrated in the second last column of Fig. 7. The neural networks without the WCMF module tend to regard background areas as salient regions, increasing the number of false positive samples, e.g., the first two images. Thus, we confirm that the WCMF module is essential.

(4) The usefulness of edge supervision. Although edge information is effective for RGB saliency detection [42], its application for RGB-D SOD has not been tested. As mentioned in Section 3.4, we added edge supervision in the second stage of the decoder. In this subsection, edge information is substituted by ground truth masks, which is represented as “w/o edge” in the following. As shown in Table 3, this replacement causes performance degradation. Visual examples are listed in the last column of Fig. 7. Compared with the proposed FCMNet, this approach blurs the boundaries of the final outputs, e.g., the right side of the third picture.

pared with the proposed FCMNet, this approach blurs the boundaries of the final outputs, e.g., the right side of the third picture.

4.4. Failure Cases

In this subsection, we present some failure cases and analyze the reasons. As shown in Fig. 8, the failure of our model is mainly caused by three kinds of reasons. First, the performance may decrease when the salient object has complex outlines, such as the first example in Fig. 8. Since the complex outlines are easily adjacent to the background, our networks can not obtain sharp boundaries. Second, when multiple objects appear in a cluttered scene, the proposed method can not obtain accurate results. For instance, the second example in Fig. 8 illustrates this case. Different objects contain various depth information, while these cues are contradictory. The neural networks will be misguided by this conflicting information. Finally, when the quality of the depth map is unsatisfactory, the proposed FCMNet may not extract the complete structure of the salient objects, as shown in the last row in Fig. 8.

5. Conclusion

In this paper, we present a two-stream encoder-decoder neural network to effectively extract and fuse the representative features in RGB-D SOD. The FACMA module is designed to automatically extract and select complementary features. In the FACMA module, we also proposed an SFCA module to preserve rich features from two modalities. The WCMF module is proposed for enhancing and fusing heterogeneous features. Extensive experiments are conducted on eight benchmark datasets. Compared with seventeen state-of-the-art methods, our model obtains competitive results on four evaluation metrics. In the future, we will focus on designing a lightweight neural network to detect RGB-D salient objects.

CRedit authorship contribution statement

Xiao Jin: Conceptualization, Methodology, Software, Writing - original draft. **Chunle Guo:** Resources. **Zhen He:** Software, Data curation. **Jing Xu:** Supervision. **Yongwei Wang:** Resources. **Yuting Su:** Resources.

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.

References

- [1] 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 Transactions on Pattern Analysis*

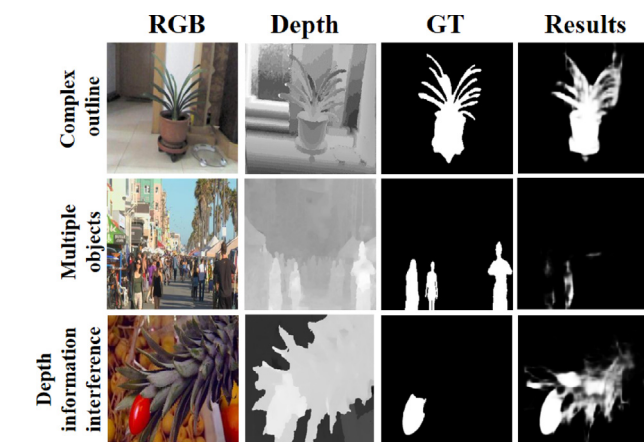


Fig. 8. Some failure cases of our FCMNet. The first row shows the salient objects with complex outlines. The second row shows the scenes with multiple salient objects. The third row shows the interference caused by depth information.

- and Machine Intelligence (2021), <https://doi.org/10.1109/TPAMI.2021.3051099>.
- [2] Y. Ji, H. Zhang, Z. Zhang, M. Liu, CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances, *Information Sciences* 546 (2021) 835–857, <https://doi.org/10.1016/j.ins.2020.09.003>.
 - [3] F. Sun, T. Kong, W. Huang, C. Tan, B. Fang, H. Liu, Feature pyramid reconfiguration with consistent loss for object detection, *IEEE Transactions on Image Processing* 28 (10) (2019) 5041–5051, <https://doi.org/10.1109/TIP.2019.2917781>.
 - [4] W. Zhang, A. Borji, Z. Wang, P. Le Callet, H. Liu, The application of visual saliency models in objective image quality assessment: A statistical evaluation, *IEEE Transactions on Neural Networks and Learning Systems* 27 (6) (2015) 1266–1278, <https://doi.org/10.1109/TNNLS.2015.2461603>.
 - [5] L. Zhang, L. Wang, W. Lin, Conjunctive patches subspace learning with side information for collaborative image retrieval, *IEEE Transactions on Image Processing* 21 (8) (2012) 3707–3720, <https://doi.org/10.1109/TIP.2012.2195014>.
 - [6] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, *IEEE Transactions on Image Processing* 19 (1) (2009) 185–198, <https://doi.org/10.1109/TIP.2009.2030969>.
 - [7] H. Fang, S. Gupta, F. Iandola, R.K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J.C. Platt, et al., From captions to visual concepts and back, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482, <https://doi.org/10.1109/CVPR.2015.7298754>.
 - [8] C. Chen, S. Li, H. Qin, A. Hao, Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis, *Pattern Recognition* 52 (2016) 410–432, <https://doi.org/10.1016/j.patcog.2015.09.033>.
 - [9] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, M.-M. Cheng, Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks, *IEEE Transactions on Neural Networks and Learning Systems* 32 (5) (2021) 2075–2089, <https://doi.org/10.1109/TNNLS.2020.2996406>.
 - [10] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, L. Shao, RGB-D salient object detection: A survey, *Computational Visual Media* 7 (1) (2021) 37–69, <https://doi.org/10.1007/s41095-020-0199-z>.
 - [11] J. Wu, W. Zhou, T. Luo, L. Yu, J. Lei, Multiscale multilevel context and multimodal fusion for RGB-D salient object detection, *Signal Processing* 178 (2021), 107766, <https://doi.org/10.1016/j.sigpro.2020.107766>.
 - [12] H.-B. Bi, Z.-Q. Liu, K. Wang, B. Dong, G. Chen, J.-Q. Ma, Towards accurate RGB-D saliency detection with complementary attention and adaptive integration, *Neurocomputing* 439 (2021) 63–74, <https://doi.org/10.1016/j.neucom.2020.12.125>.
 - [13] Z. Qin, P. Zhang, F. Wu, X. Li, FcaNet: Frequency Channel Attention Networks, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 783–792.
 - [14] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (8) (2020) 2011–2023, <https://doi.org/10.1109/TPAMI.2019.2913372>.
 - [15] A.G. Roy, N. Navab, C. Wachinger, Concurrent spatial and channel -squeeze and excitation' in fully convolutional networks, in: *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*, 2018, pp. 421–429, https://doi.org/10.1007/978-3-030-00928-1_48.
 - [16] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19, https://doi.org/10.1007/978-3-030-01234-2_1.
 - [17] J. Park, S. Woo, J.-Y. Lee, I.S. Kweon, BAM: Bottleneck attention module, *arXiv preprint arXiv:1807.06514* (2018).
 - [18] Z. Gao, J. Xie, Q. Wang, P. Li, Global second-order pooling convolutional networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3024–3033, <https://doi.org/10.1109/CVPR.2019.00314>.
 - [19] N. Liu, L. Li, W. Zhao, J. Han, L. Shao, Instance-level relative saliency ranking with graph reasoning, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), <https://doi.org/10.1109/TPAMI.2021.3107872>, 1–1.
 - [20] N. Liu, W. Zhao, L. Shao, J. Han, SCG: Saliency and Contour Guided Salient Instance Segmentation, *IEEE Transactions on Image Processing* 30 (2021) 5862–5874, <https://doi.org/10.1109/TIP.2021.3088282>.
 - [21] N. Liu, W. Zhao, D. Zhang, J. Han, L. Shao, Light field saliency detection with dual local graph learning and reciprocal guidance, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 4712–4721.
 - [22] N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 4722–4732.
 - [23] N. Liu, J. Han, M.-H. Yang, PiCANet: Pixel-Wise Contextual Attention Learning for Accurate Saliency Detection, *IEEE Transactions on Image Processing* 29 (2020) 6438–6451, <https://doi.org/10.1109/TIP.2020.2988568>.
 - [24] N. Liu, N. Zhang, J. Han, Learning Selective Self-Mutual Attention for RGB-D Saliency Detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13756–13765.
 - [25] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 454–461, <https://doi.org/10.1109/CVPR.2012.6247708>.
 - [26] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: a benchmark and algorithms, in: *Proceedings of the European Conference on Computer Vision*, 2014, pp. 92–109, https://doi.org/10.1007/978-3-319-10578-9_7.
 - [27] D. Feng, N. Barnes, S. You, C. McCarthy, Local Background Enclosure for RGB-D Salient Object Detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2343–2350, <https://doi.org/10.1109/CVPR.2016.257>.
 - [28] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, T. Ren, Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning, *IEEE Transactions on Image Processing* 26 (9) (2017) 4204–4216, <https://doi.org/10.1109/TIP.2017.2711277>.
 - [29] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, S. Kwong, Going From RGB to RGBD Saliency: A Depth-Guided Transformation Model, *IEEE Transactions on Cybernetics* 50 (8) (2020) 3627–3639, <https://doi.org/10.1109/TCYB.2019.2932005>.
 - [30] X. Zhu, Y. Li, H. Fu, X. Fan, Y. Shi, J. Lei, RGB-D salient object detection via cross-modal joint feature extraction and low-bound fusion loss, *Neurocomputing* 453 (2021) 623–635, <https://doi.org/10.1016/j.neucom.2020.05.110>.
 - [31] 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, <https://doi.org/10.1016/j.neucom.2020.01.045>.
 - [32] Z. Huang, H.-X. Chen, T. Zhou, Y.-Z. Yang, B.-Y. Liu, Multi-level cross-modal interaction network for RGB-D salient object detection, *Neurocomputing* 452 (2021) 200–211, <https://doi.org/10.1016/j.neucom.2021.04.053>.
 - [33] 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, <https://doi.org/10.1016/j.patcog.2018.08.007>.
 - [34] 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, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936, <https://doi.org/10.1109/CVPR.2019.00405>.
 - [35] Y. Piao, W. Ji, J. Li, M. Zhang, H. Lu, Depth-induced multi-scale recurrent attention network for saliency detection, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7254–7263, <https://doi.org/10.1109/ICCV.2019.00735>.
 - [36] C. Li, R. Cong, Y. Piao, Q. Xu, C.C. Loy, RGB-D salient object detection with cross-modality modulation and selection, in: *Proceedings of the European Conference on Computer Vision*, 2020, pp. 225–241, https://doi.org/10.1007/978-3-030-58598-3_14.
 - [37] M. Zhang, W. Ren, Y. Piao, Z. Rong, H. Lu, Select, supplement and focus for RGB-D saliency detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3472–3481, <https://doi.org/10.1109/CVPR42600.2020.00353>.
 - [38] Y. Piao, Z. Rong, M. Zhang, W. Ren, H. Lu, A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9060–9069, <https://doi.org/10.1109/CVPR42600.2020.00908>.
 - [39] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F.S. Saleh, T. Zhang, N. Barnes, UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8582–8591.
 - [40] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F.E. Alsaadi, A survey of deep neural network architectures and their applications, *Neurocomputing* 234 (2017) 11–26, <https://doi.org/10.1016/j.neucom.2016.12.038>.
 - [41] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: *Proceedings of the European Conference on Computer Vision*, 2018, pp. 833–851, https://doi.org/10.1007/978-3-030-01234-2_49.
 - [42] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, M.-M. Cheng, EGNet: Edge guidance network for salient object detection, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8779–8788, <https://doi.org/10.1109/ICCV.2019.00887>.
 - [43] Q. Zhang, T. Xiao, N. Huang, D. Zhang, J. Han, Revisiting Feature Fusion for RGB-T Salient Object Detection, *IEEE Transactions on Circuits and Systems for Video Technology* 31 (5) (2021) 1804–1818, <https://doi.org/10.1109/TCSVT.2020.3014663>.
 - [44] H. Chen, Y. Li, Three-Stream Attention-Aware Network for RGB-D Salient Object Detection, *IEEE Transactions on Image Processing* 28 (6) (2019) 2825–2835, <https://doi.org/10.1109/TIP.2019.2891104>.
 - [45] H. Chen, Y. Li, Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.
 - [46] Q. Zhang, T. Xiao, N. Huang, D. Zhang, J. Han, Revisiting Feature Fusion for RGB-T Salient Object Detection, *IEEE Transactions on Circuits and Systems for Video Technology* 31 (5) (2021) 1804–1818, <https://doi.org/10.1109/TCSVT.2020.3014663>.
 - [47] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, in: *Proceedings of the International Conference on Internet Multimedia Computing and Service*, 2014, pp. 23–27, <https://doi.org/10.1145/2632856.2632866>.
 - [48] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in: *Proceedings of the IEEE International Conference on Image Processing*, 2014, pp. 1115–1119, <https://doi.org/10.1109/ICIP.2014.7025222>.

- [49] C. Zhu, G. Li, A three-pathway psychobiological framework of salient object detection using stereoscopic technology, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 3008–3014, <https://doi.org/10.1109/ICCVW.2017.355>.
- [50] N. Li, J. Ye, Y. Ji, H. Ling, J. Yu, Saliency detection on light field, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2806–2813, <https://doi.org/10.1109/CVPR.2014.359>.
- [51] N. Liu, N. Zhang, L. Shao, J. Han, Learning Selective Mutual Attention and Contrast for RGB-D Saliency Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1–1, <https://doi.org/10.1109/TPAMI.2021.3122139>.
- [52] 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 IEEE International Conference on Computer Vision, 2017, pp. 4548–4557, <https://doi.org/10.1109/ICCV.2017.487>.
- [53] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1597–1604, <https://doi.org/10.1109/CVPR.2009.5206596>.
- [54] A. Borji, M.-M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Transactions on Image Processing 24 (12) (2015) 5706–5722, <https://doi.org/10.1109/TIP.2015.2487833>.
- [55] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2018, pp. 698–704, <https://doi.org/10.5555/3304415.3304515>.
- [56] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, Q. Huang, ASIF-Net: Attention Steered Interweave Fusion Network for RGB-D Salient Object Detection, IEEE Transactions on Cybernetics 51 (1) (2021) 88–100, <https://doi.org/10.1109/TCYB.2020.2969255>.
- [57] X. Wang, S. Li, C. Chen, A. Hao, H. Qin, Depth quality-aware selective saliency fusion for RGB-D image salient object detection, Neurocomputing 432 (2021) 44–56, <https://doi.org/10.1016/j.neucom.2020.12.071>.
- [58] H. Chen, Y. Li, Three-Stream Attention-Aware Network for RGB-D Salient Object Detection, IEEE Transactions on Image Processing 28 (6) (2019) 2825–2835, <https://doi.org/10.1109/TIP.2019.2891104>.
- [59] H. Chen, Y. Li, Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3051–3060, <https://doi.org/10.1109/CVPR.2018.00322>.
- [60] J. Han, H. Chen, N. Liu, C. Yan, X. Li, CNNs-Based RGB-D Saliency Detection via Cross-View Transfer and Multiview Fusion, IEEE Transactions on Cybernetics 48 (11) (2018) 3171–3183, <https://doi.org/10.1109/TCYB.2017.2761775>.
- [61] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, Q. Yang, RGBD Salient Object Detection via Deep Fusion, IEEE Transactions on Image Processing 26 (5) (2017) 2274–2285, <https://doi.org/10.1109/TIP.2017.2682981>.
- [62] C. Zhu, G. Li, W. Wang, R. Wang, An innovative salient object detection using center-dark channel prior, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1509–1515, <https://doi.org/10.1109/ICCVW.2017.178>.
- [63] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, C. Hou, Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion, IEEE Signal Processing Letters 23 (6) (2016) 819–823, <https://doi.org/10.1109/LSP.2016.2557347>.
- [64] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
- [65] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, <https://doi.org/10.1109/CVPR.2009.5206848>.
- [66] S. Ruder, An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747 (2016).



Xiao Jin received the B.S. degree and Ph.D. degree from Tianjin University, Tianjin, China, both in electronic information engineering. He is currently an assistant professor with the College of Artificial Intelligence from Nankai University, Tianjin, China. From 2017 to 2018, he was a visiting student with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada. His current research interests include computer vision and multimedia computing.



Chunle Guo received the Ph.D. degree from Tianjin University, China. He is currently a Postdoctoral Research Fellow at Nankai University. His research interests lie in image processing, computer vision, and deep learning.



Zhen He received the B.S. degree from Nantong University in 2018. He is currently pursuing the M.S. degree with the College of Artificial Intelligence from Nankai University, Tianjin, China. His research interests include computer vision and multimedia computing.



Jing Xu received the Ph.D. degree from Nankai University, in 2003. She is currently a Professor with the College of Artificial Intelligence, Nankai University. Her current research interests include intelligent software engineering and medical data analysis. She received the Second Prize of the Tianjin Science and Technology Progress Award, in 2017 and 2018.



Yongwei Wang received the B.S. and M.S. degree from Northwestern Polytechnical University, China, in electronic information engineering. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of British Columbia, Canada. His current research interests include multimedia content analysis and security.



Yuting Su received the M.S. and Ph.D. degrees in electrical engineering from Tianjin University, Tianjin, China, in 1995, 1998, and 2001 respectively. He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. He was a Visiting Scholar with Case Western Reserve University, Cleveland, OH, USA, from 2009 to 2010. His research interests include computer vision and multimedia computing.