Computer Science and Information Systems 21(3):947–969 https://doi.org/10.2298/CSIS231206022X

ECW-EGNet: Exploring Cross-Modal Weighting and Edge-Guided Decoder Network for RGB-D Salient Object Detection

Chenxing Xia[1,2,3], Feng Yang[1,⋆], Songsong Duan[4], Xiuju Gao[5], Bin Ge[1], Kuan-Ching Li[6], Xianjin Fang[1,7], Yan Zhang[8], and Ke Yang[2,⋆]

[1] College of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, 232001, China
2021201237@aust.edu.cn
[2] The First Affiliated Hospital of Anhui University of Science and Technology, (Huainan First People's Hospital), China
cxxia@aust.edu.cn
[3] Anhui Purvar Bigdata Technology Co. Ltd, Huainan, 232001, China
[4] State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
[5] College of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, Anhui, China
[6] Department of Computer Science and Information Engineering, Providence University, Taiwan
[7] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[8] The School of Electronics and Information Engineering, Anhui University, Hefei, Anhui, China

Abstract. Existing RGB-D salient object detection (SOD) techniques concentrate on combining data from multiple modalities (e.g., depth and RGB) and extracting multi-scale data for improved saliency reasoning. However, they frequently perform poorly as a factor of the drawbacks of low-quality depth maps and the lack of correlation between the extracted multi-scale data. In this paper, we propose a Exploring Cross-Modal Weighting and Edge-Guided Decoder Network (ECW-EGNet) for RGB-D SOD, which includes three prominent components. Firstly, we deploy a Cross-Modality Weighting Fusion (CMWF) module that utilizes Channel-Spatial Attention Feature Enhancement (CSAE) mechanism and Depth-Quality Assessment (DQA) mechanism to achieve the cross-modal feature interaction. The former parallels channel attention and spatial attention enhances the features of extracted RGB streams and depth streams while the latter assesses the depth-quality reduces the detrimental influence of the low-quality depth maps during the cross-modal fusion. Then, in order to effectively integrate multi-scale features for high-level and produce salient objects with precise locations, we construct a Bi-directional Scale-Correlation Convolution (BSCC) module in a bi-directional structure. Finally, we construct an Edge-Guided (EG) decoder that uses the edge detection operator to obtain edge masks to guide the enhancement of salient map edge details. The comprehensive experiments on five benchmark RGB-D SOD datasets demonstrate that the proposed ECW-EGNet outperforms 21 state-of-the-art (SOTA) saliency detectors in four widely used evaluation metrics.

Keywords: cross-modality fusion, depth-quality, edge-guided, RGB-D images, salient object detection.

⋆ Corresponding authors

1. Introduction

Salient object detection (SOD), a crucial component of computer vision, tries to mimic the visual learning mechanism of the human body to react fast to visual stimuli and produce visually appealing or fascinating objects or regions by exploring and segmenting particular objects. As a valuable preprocessing tool, it has been used extensively in various computer vision applications, including semantic segmentation [1], image retrieval [2] , visual tracking [3], and remote sensing image segmentation [4]. The technique can perform a series of image processing operations and speed up the processing of data by intelligently, effectively, and precisely recognizing the salient object regions. It can also effectively distribute the limited computing resources to some extent.

In traditional RGB-based SOD, the color and texture cues captured by RGB images play a crucial role in identifying salient objects. However, their performance suffers when faced with difficult situations such complex backgrounds and poor illumination. As a result of the development of sensors like smartphones and Microsoft Kinect depth camera, it is now possible to quickly gather both RGB maps and equivalent depth maps from a scene. The spatial information and geometry contained in the depth map can be used to offset the shortcomings of the RGB image, which provides detailed information. As a result, it is a reasonable choice to combine RGB images and depth maps for SOD tasks (called RGB-D SOD), which can handle more complicated scenarios and satisfy the needs of advanced detection. However, directly integrating RGB images and depth images is not a good countermeasure and introduces unique challenges. These challenges include: (1) Depth noise: Depth maps acquired from depth sensors can be affected by noise and inaccuracies, which can impact the quality and accuracy of saliency maps. (2) Complementary fusion: Integrating RGB and depth information to capture complementary cues and avoid redundancy is crucial. Developing effective fusion strategies that fully utilize the strengths of each modality is a challenge.

The depth map is generally used as prior knowledge to assist handcrafted features for SOD in the early stages of RGB-D SOD approaches. Unfortunately, early handcrafted feature techniques were really crude, frequently unable to represent data with rich high-level semantics, and their performance was subpar. In past decades, feature representation in convolutional neural networks (CNNs) has grown functional because CNNs can acquire features from the geometry, color, and spatial information of images, causing RGB-D SOD technology to pay greater attention to deep learning. Existing approaches focus on how to design effective ”interaction/fusion” modules that bridge the gap between the two modes. Wei *et al*. [5] sufficiently melded the bidirectional attention interaction module achieves bidirectional interaction between cross-modality features. Wang *et al*. [6] converted the original 4-dimensional RGB-D cycle into DGB, RDB, and RGD to fuse RGB and depth before extracting depth features, which helps achieve the best channel complementary fusion state between RGB and depth. Chen *et al*. [7] adaptively fusion of multi-modal information makes use of complimentary RGB and depth cues to address the issue of inaccurate depth maps. Xia *et al*. [8] introduced a bi-directional interactive architecture in order to improve RGB and depth features through a circular interaction. Bi *et al*. [9] proposed a feedback mechanism that receives two input features from RGB and depth branches to exchange would exchange the new information into existing interactive information. Zeng *et al*. [10] propose a compensated attention feature fusion and hierarchical multiplication decoder network. Although these models have achieved remarkable

results in exploring feature fusion between modalities, they ignore the possible negative effects caused by the introducing low-quality depth maps in the fusion process.

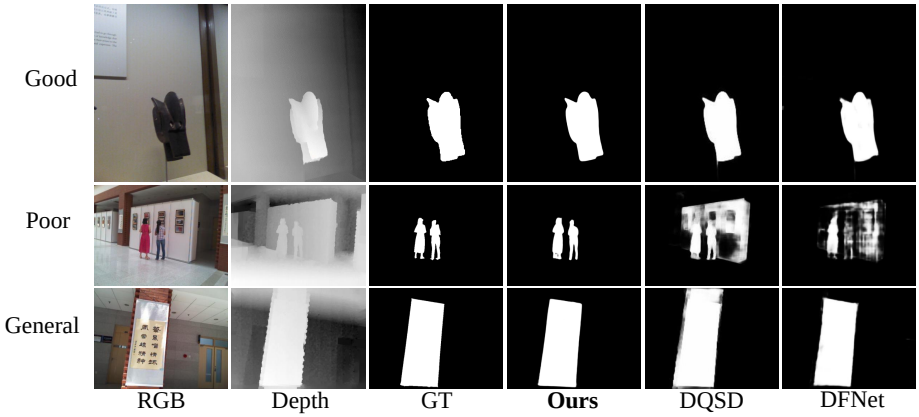

Fig. 1. Visual examples of our method, DQSD [11] and DFNet [12] of depth maps with Good, Poor, and General (top-to-bottom)

In order to avoid the contamination of low-depth maps, some works [11,13,14] develop depth awareness, a guidance method of fusing cross-modal data. On the other hand, some researchers [12,15] enhance depth maps and suppress noise from depth maps using feature enhancement modules. In order to improve the accuracy, a growing number of techniques [16,17] are devoted to presenting elaborate feature fusion modules, which include well-designed alternate interactions between features to filter out noise in RGB and depth features with the help of other modal data. Furthermore, several methods [18,19] exploit estimates supplementary meaning depth maps to lessen the impact of low-quality depth maps. However, there are some problems with these approaches: (1) Feature enhancement and elaborate feature fusion mainly focus on complementary and ignore the specificity between different modalities; (2) Depth-aware and depth estimation methods undoubtedly increase the computational cost by estimating the depth map; (3) The performance of the SOD is significantly impacted by effective mechanisms for assessing the quality of depth maps. Some visual examples are shown in Figure 1, where the depth map quality is classified as good, poor, and general. In these cases, features extracted from low-quality input images reduce the discriminative ability of fused features. Therefore, it is important to consider the quality of depth maps in the RGB-D SOD task.

Furthermore, it is important to note that SOD is fundamentally a pixel-wise dense prediction work, which utilizes the categorization of pixels on feature maps by relying on high-resolution and multi-scale features. However, the latest advances are mainly based on atrous convolution techniques, which enhance high-level features through the integration of semantic segmentation modules. One of the most well-known modules is the ASPP and its variant versions [20,21,22,23], which adjust the receptive field size and enhance the expressive power of the feature map by controlling the atrous rate without introducing additional parameters. Although these models have achieved remarkable results, they

still have two limitations: (1) ***The loss of semantic information***. The calculation of atrous convolutions is similarly to a checkerboard format, where the convolution kernel only focuses on certain positions of the pixels in the feature map during computation, while other positions are replaced by far "zero elements". This means that the output results are determined only by pixels in a few positions, which may lead to the loss of semantic information. (2) ***The lack of information correlation***. These methods capture a wide range of contextual information by varying the atrous convolution. However, in practical applications, the feature representations across different scales often exhibit correlations. Separate multi-scale pooling operations on the outputs of the deep network do not adequately consider this correlation, which may also be detrimental to the generation of high-quality salient maps.

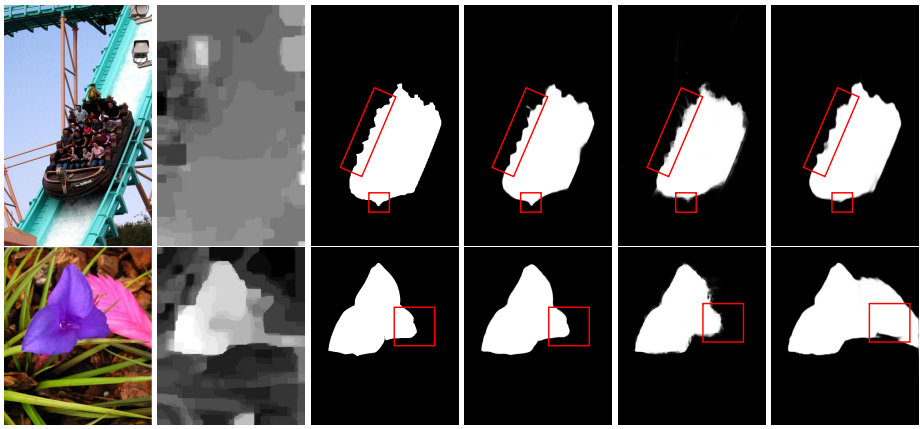

Fig. 2. Compare the results of our model with JL-DCF [24] and SMEG [25]. The red anchor box represents the comparison area

Moreover, edge features refer to the information around objects in an image, which can provide important clues about object shape and structure. In SOD, these edge features are crucial for distinguishing salient objects from the background, as salient objects typically have distinct contours and edge features that differ from the surrounding background. By effectively extracting and analyzing the edge features of an image, salient objects can be more accurately identified and separated them from the background. However, most of the past research has primarily focused on the structural integrity of SOD results rather than edge quality [26]. As a result, many existing approaches suffer from poor edge quality in their output saliency maps. As shown in the first row of Figure 2, when performing SOD tasks in simple scenes, JL-DCF [24] and SMEG [25] can effectively extract objects from the background. However, the edge details captured by these models appear slightly rougher compared to those produced by the proposed ECW-EGNet method. This observation highlights the need for accurate edge localization and the incorporation of

edge attention features to optimize the performance of RGB-D SOD algorithms, thereby mitigating the issue of edge blurring, as shown in the red box in the figure.

To solve the above problems, a Exploring Cross-Modal Weighting and Edge-Guided Decoder Network (ECW-EGNet) is proposed for RGB-D SOD, which is equipped with a cross-modal weighting fusion (CMWF) module, a bi-directional scale-correlation convolution (BSCC) module, and an edge-guided decoder. Specifically, the CMWF module is designed to enhance the complementary between the cross-modal features and better integrate the features, where Channel-Spatial Attention Feature Enhancement (CSAE) mechanism and depth-quality assessment (DQA) mechanism are embedded to enhance the discriminant feature and adaptively weights and refines the RGB and depth branches, respectively. In addition, to exploit the property of fused multi-scale features, the BSCC module is constructed to better capture scale-correlation information and integrate multi-scale features of high-level features by employing depth-wise separable convolutions to compensate for the atrous convolution, where each feature branch is connected in series to fill the holes and further improve the quality of saliency maps. Finally, to enhance the edge information, an Edge-Guided (EG) decoder is proposed to synthesize more accurate boundary masks by means of an edge detection operator, thus guiding the decoder to achieve more accurate image segmentation.

The main contributions of this paper include the following:

- We suggest a Exploring Cross-Modal Weighting Fusion and Edge-Guided Decoder Network (ECW-EGNet) for RGB-D SOD that not only assesses the quality of the depth map but also utilizes the feature attributes of different scales and edge-enhancement features to improve the efficiency of generating high-quality saliency maps.
- We propose a Cross-Modal Weighting Fusion (CMWF) module to facilitate the learning of complementary information across modalities, in which a Channel-Spatial Attention Feature Enhancement (CSAE) mechanism is employed to enhance the features of extracted RGB streams and depth streams, and a Depth-Quality Assessment (DQA) mechanism is embedded to assess the quality of the feature of the depth map by calculating the weighting factors to reduce the interference problem caused by the low-quality of the depth map.
- We construct a Bi-directional Scale-Correlation Convolution (BSCC) module based on a well-designed bi-directional structure to capture the complementary and correlated information of high-level features from different scales and reduce the loss of local information.
- We propose an Edge-Guided (EG) decoder that guides the upgrading of salient object edges by using an edge detection operator to synthesize more accurate edge masks.

2. Materials and Methods

2.1. Network Overview

The overall framework of the proposed ECW-EGNet is shown in Figure 3. We adopt the Swin Transformer [27] as the backbone of a two-stream architecture to acquire multi-level feature representation via RGB and depth maps. Swin Transformer has a powerful ability to model far range dependencies and capture global context. Formally, the generated from the RGB and depth streams are named as $\{f_i^r\}_{i=1}^4$ and $\left\{f_i^d\right\}_{i=1}^4$, respectively. The

discriminant representation of the CMWF with enhanced fusion features is proposed to obtain multi-levels of fusion features by integrating RGB and depth features. Notably, F_3^f and F_4^f are sent to a bi-directional scale-correlation convolution (BSCC) module to gather high-level semantic features and enhance the inference of the result. For the decoder, we use the edge detection operator (i.e., Sobel [28]) to generate edge-enhanced features for subsequent edge-guided (EG) decoders to retain a clearer edge

2.2. Cross-Modality Weighting Fusion (CMWF) module

The interaction between RGB features and depth features is essential for RGB-D SOD. However, if low-quality depth maps are chosen to be used in the cross-modal fusion process, interference and non-salient objects may be included, and resulting in suboptimal performance. CMWF is deployed to leverage the unique properties of RGB and depth information in order to enhance the fusion process. By assigning appropriate weights to these modalities, the mechanism aims to effectively capture and combine the complementary information they provide. As illustrated in Figure 4, CMWF consists of three main components: channel-spatial attention feature enhancement, depth-quality assessment mechanism and depth-weighted cross-modality fusion.

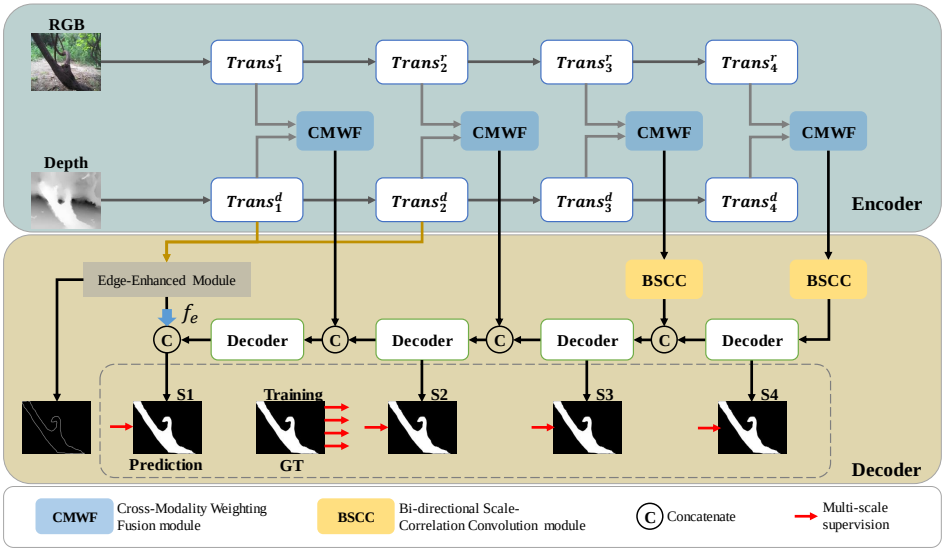

Fig. 3. Architecture of the proposed method. The feature encoding, cross-modal weighting fusion, and saliency inference are depicted as three essential processes in the part

Channel-Spatial Attention Feature Enhancement(CSAE). It is increasingly acknowledged that RGB images and depth maps are essentially two distinct kinds of information representations in RGB-D SOD. RGB information contains rich color and texture details, which are particularly useful for capturing surface appearance and object bound-

aries. On the other hand, depth information provides valuable depth cues and shape information, which are crucial for understanding object geometry and spatial relationships. Therefore, the features of each modality hold different importance in the channel. A Simple fusion of these features may introduce noise, thereby requiring feature enhancement to increase the discrimination between modalities, facilitating the subsequent feature assessment. Attention mechanisms have been introduced to the computer vision system, which is motivated by the human visual system. Across the previous years, it has seen an increase in the significance of attention mechanisms in computer vision. In general, there are three distinct categories of attention mechanisms: channel attention mechanisms, spatial attention mechanisms, and combination mechanisms that combine the former two classes. Different from previous works [29,30], which combined channel and spatial attention mechanisms serially, we use a parallel approach to infer attention mapping for each branch (spatial and channel). As illustrated in Figure 4, this parallel channel-spatial attention could simultaneously concentrate on "what" and "where", which captures the importance of different feature channels, emphasizes common salient objects, and removes unnecessary noise.

Specifically, we calculate channel attention maps and spatial attention maps of RGB feature f_i^r and depth feature f_i^d, which are defined as follows:

$$CA_i^c = Sigmoid(Conv_1(GAP(f_i^c))), \tag{1}$$

$$SA_i^c = Sigmoid(Conv_3(CGAP(f_i^c))), \tag{2}$$

where $c \in (r, d)$, $i \in (1, 2, 3, 4)$, CA_i^c and SA_i^c present the channels and spatial attention maps at the i-th level, respectively. $GAP(\cdot)$ represents the global average pooling operation, $CGAP(\cdot)$ means the global average pooling operation along channel direction. $Conv_k(\cdot)$ represents the convolution operation with the kernel size $k \times k (k = 1, 3)$, and $Sigmoid(\cdot)$ represents the sigmoid activation function. Next, the original features with the attention are combined to create the enhanced features, which are described as follows:

$$F_i^c = f_i^c \times (CA_i^c \times SA_i^c), \tag{3}$$

where $c \in (r, d)$, F_i^c is the feature enhanced by channel-spatial attention, and $\times$ represents the operations of element-wise multiplication.

Depth-Quality Assessment (DQA) Mechanism. RGB images and depth maps are essential components of RGB-D SOD, and the fused features often serve as the key breakthroughs in RGB-D SOD performance during cross-modal data fusion. However, low-quality depth images resulting from factors such as dark lighting, random noise, and human interference can lead to information loss. Directing the fusion of features extracted from RGB and depth streams without discrimination will lead to adverse outcomes. Therefore, it is necessary to develop quality assessment mechanisms and feature screening for depth maps. Inspired by [11,31,32] that high-quality depth map should be at the same level of quality as the RGB image and have similar distinctive features, we design a depth-quality assessment (DQA) mechanism to compute depth weighting factors. As shown in Figure 4, we weight the enhanced feature F_i^d to prevent excessive noise

introduction. Specifically, we assess the similarity between the depth feature-level attention map and the RGB feature-level attention map to reduce the importance of these depth maps. The quality assessment mechanism is a crucial part of CMWF, which minimizes the negative effects of low-quality depth maps on SOD.

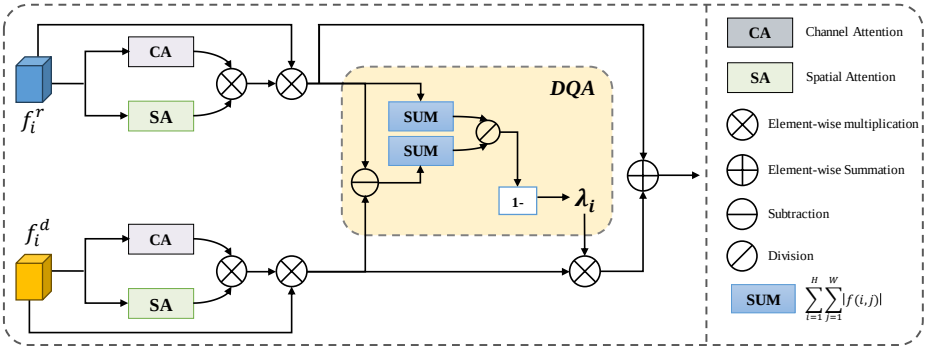

Fig. 4. The structure of the cross-modality weighting Fusion (CMWF) module."1-" means 1 subtract the input of the equation

Specifically, calculating the difference between the enhanced RGB features and the depth features is a subtraction operation, which generates the absolute value difference between the features of the two modalities. Then, the resulting difference of the absolute value of the RGB feature pixel value is utilited to get the i-th level weighting factor λ_i, which can be defined as follows:

$$S_i^{rd} = Conv_1(F_i^r) - Conv_1(F_i^d), \tag{4}$$

$$\lambda_i = 1 - \frac{SUM(S_i^{rd})}{SUM(F_i^r)}, \tag{5}$$

where $SUM(f) = \sum_{i=1}^{H}\sum_{i=1}^{H}|f(i,j)|$, $|\cdot|$ represent the absolute value operation, H and W are the height and width of feature f, respectively.

Depth-weighted Cross-Modality Fusion. This process is shown in Figure 4, we combine RGB and depth features in a weighting factors manner to obtain fusion features F_i^f, The process can be expressed as:

$$F_i^f = F_i^r + F_i^d \times \lambda_i, \tag{6}$$

this module uses the depth information to weigh the contribution of depth modality during the fusion process, thus improving accuracy and reducing noise.

2.3. Bi-directional Scale-Correlation Convolution (BSCC)

Generally, low-level features focus on the fine-grained details of the learning object, while high-level features provide more semantic information. Several approaches [21,33,34,35]

often use atrous convolution with different dilation rates to aggregate multi-scale features and inference context information. However, this approach leads to the loss of some details and local semantic information. Moreover, in the process of multi-scale information fusion, features across different scales lack correlation, which may not be conducive to generating high-quality saliency maps. To leverage the advantages of high-level features and long-distance information correlation, we design a bi-directional scale-correlation convolution (BSCC) module. The module extracts features step-by-step via top-down and bottom-up bi-directional paths. As shown in Figure 5, the BSCC contains four branches and connects them in series to fill in the holes. Each branch passes through a series of fine-grained convolutional operations to extract higher-level semantic information.

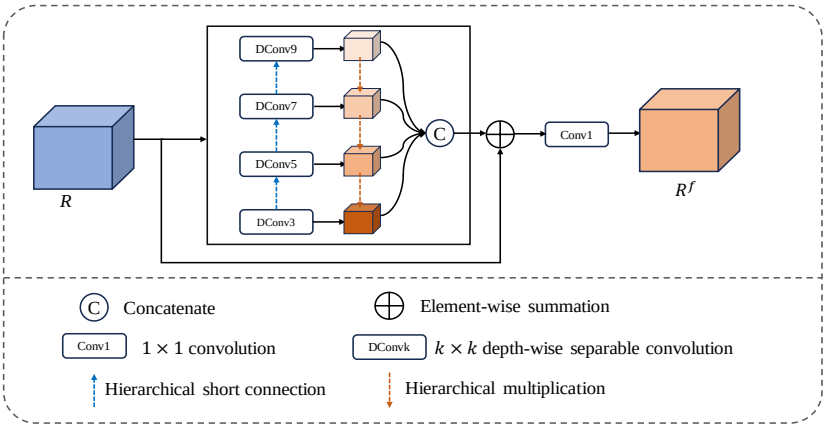

Fig. 5. The structure of the BSCC for multi-scale information fusion

Specifically, hierarchical short connections are first used to successively obtain four-scale features from the depth-wise separable convolution layers of the convolution kernel $(3, 5, 7,$ and $9)$. In addition, through hierarchical multiplication, the proposed bi-directional scale-correlation convolution (BSCC) module enhances both high-level features and multi-scale feature learning. Suppose the input feature is denoted R:

$$\begin{cases} R_1 = DConv_3(R) + R \\ R_i = DConv_{2\times i+1}(R_{i-1}) + R, i \in (2, 3, 4), \\ \tilde{R}_i = R_{i+1} \times R_i, i \in (1, 2, 3) \end{cases} \tag{7}$$

where $DConv_3$ represent a 3×3 depth-wise separable convolution, and $DConv_{2\times i+1}$ represent depth-wise separable convolution with a convolution kernel of $2 \times i + 1$. Finally, we cascade the obtained multi-scale features by concatenation, and transform the feature map through a 1×1 convolutional layer. In addition, residual concatenation is designed inside the BSCC to prevent the loss of high-level semantic information, which can be expressed as:

$$R^f = Conv_1(Cat(\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, R_4)) + R. \tag{8}$$

BSCC learns scale-correlation features efficiently by increasing the receptive field using progressive scale features, thereby greatly improving the ability to effectively infer salient objects from complex real-world scenes.

2.4. Edge-Guided (EG) decoder

This section focuses on the edge-guided decoder, which includes feature optimisation based on the edge-enhanced mechanism and information fusion strategies based on edge-guided fusion.

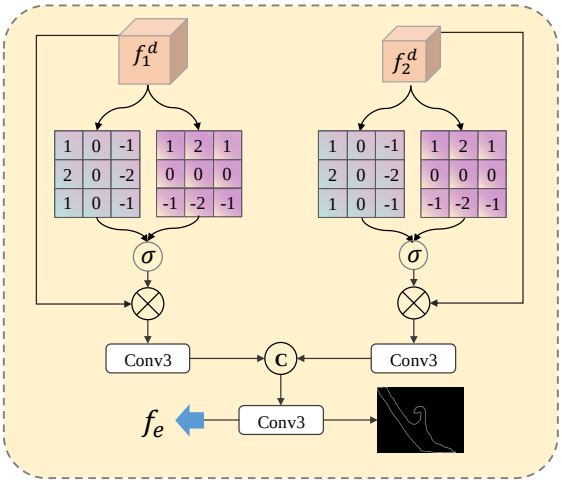

Fig. 6. The architecture of the Edge-Enhanced Module

Edge-Enhanced Mechanism. Edges are one of the important features in saliency object detection, and typically RGB image based edge enhancement methods are not as effective as they could be, mainly because RGB image low-level features only contain simple representations of surface information such as colors and textures, which are susceptible to interference when the background texture of some application scenes is complex. In contrast, depth maps are able to directly capture the difference in distance between the surface of an object and the background, better preserving detailed information such as the edge of the object. Figure 6 shows how EEM extracts edge information by using depth low-level features f_1^d and f_2^d to filter out unimportant edge information. To accomplish this, we build the gradient map using the Sobel operator [28] in both the horizontal G_x and vertical G_y direction. Convolution procedures are performed with two fixed $3\times$ 3 parameters with a stride of 1. One definition of these convolutions is:

$$M_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \qquad M_y = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \\ -1 & -2 & -1 \end{bmatrix}. \tag{9}$$

The two convolutions described above are then applied to the input feature map to create the gradient maps I_x andI_y. The edge-enhanced feature map f_e is produced by fusing these gradient maps with the input feature map after normalizing them with the sigmoid function:

$$f_e = f_d \times Sigmoid(Cat(I_x, I_y)), \tag{10}$$

where $Cat(\cdot)$ denotes concatenation along the channel dimension. Next, we fuse the edge-enhanced feature maps of f_1^d and f_2^d. Frist, the feature map of f_2^d is first applied to a 1×1 convolution operation, and a feature map with the same size as f_1^d is generated using bilinear upsampling. Then, to equalize the number of channels between the two feature maps, a 3×3 convolution is performed on each of the two feature maps. The final edge feature map f_e is created by concatenating the two feature maps along the channel dimension and applying two layers of convolution, with the result being supervised by the ground truth edge map.

Edge-Guided Fusion. Low-level features can provide more detailed and localized information that can be used to locate small objects well, while high-level features can capture global information and enable the detection of large objects. Inspired by the decoding idea widely used in the Unet framework [36]. According to [36], the edge-guided decoder is designed to combine the obtained multi-scale cross-modal weighting fusion features with edge-enhanced features in a progressive manner to make full use of the multi-scale features. As shown in Figure 3, multi-scale feature fusion results are obtained by channel concatenating using cross-modal weighting fusion features and edge-enhanced features as input. The methods are as follows:

$$\overline{f}_i^c = \begin{cases} Cat(F_i^f, Up(Conv_3(F_{i+1}^f))), i = 1, 2, 3 \\ F_i^f. \qquad\qquad\qquad\qquad\quad i = 4 \end{cases} \tag{11}$$

Finally, the edge features from the edge-enhanced module are combined with the fused features to generate the edge-guided salient map *S*.

$$S = Up(Cat(\overline{f}_i^c, f_e)). \tag{12}$$

2.5. Loss Function

As shown in Figure 2, saliency maps of intermediate predictions are produced during the network testing phase in each decoder block, denoted as *S*(*t*). The ground truth *GT* supervises each *S*(*t*). We adopt the Binary Cross-Entropy (BCE) loss function and the Intersection-Over-Union (IOU) loss function during the supervised training procedure to enhance the content representation. The total loss function, denoted *TLoss*, is defined as follows:

$$TLoss = \sum_{t=1}^{5} L_{bce}^{(t)}(up(S(t)), GT) + L_{iou}^{(t)}(up(S(t)), GT), \tag{13}$$

where *GT* represents the ground truth, $up(\cdot, \cdot)$ represents bilinear upsampling, and $S(t)$ is upsampled to the same resolution as GT; $L_{bce}^{(t)}(\cdot, \cdot)$ represents BCE loss, $L_{iou}^{(t)}(\cdot, \cdot)$ represents the IOU loss, and its calculation formula is defined as:

$$L_{bce} = -y \times \lg(\hat{y}) + (1 - y)\lg(1 - \hat{y}), \tag{14}$$

$$L_{iou} = \frac{|A \cap B|}{|A \cup B|}, \tag{15}$$

where y represents the analogy to which it belongs, $\hat{y}$ represents the prediction map; A represents the $S(t)$ area and B represents the GT area.

3. Experiments

3.1. Datasets

We evaluate the performance of the proposed model on five popular RGB-D SOD benchmarks, including RGBD135 dataset [37], NLPR dataset [38], NJU2K dataset [39], SSD dataset [40], and STERE [41] dataset. RGBD135 dataset [37] consists of 135 indoor RGB-D images obtained by Kinect. NLPR dataset [38] is a collection of 1000 RGB images and depth maps pairs that were acquired using Microsoft Kinect and include both indoor and outdoor scenes. NJU2K dataset [39] is a collection of 1985 RGB images and matching depth maps gathered from stereoscopic photographs, the internet, and 3D movies. The SSD [40] is made up of 80 RGB-D images, each of which is part of stereoscopic film production. The STERE [41] integrates 1000 pairs of binocular images collected from the Internet and is the first stereo image dataset in this field. For unbiased algorithm evaluation and comparison, following [42,43], the same training dataset comprising 1485 RGB-D images from the NJU2K dataset and 700 RGB-D images from the NLPR dataset was used. The remaining images were used for subsequent testing.

3.2. Evaluation Metrics

To evaluate the performance of the proposed approach and other approaches, we select four common evaluation metrics, including Precision-Recall (PR) curve, adaptive E-measure (E_{ξ}^{adp}) [44], adaptive F-measure (F_{β}^{adp}) [45], S-measure (S_{α}) [46], and Mean Absolute Error (MAE) [47].

The Precision and Recall of a saliency map *S* and a ground truth *G* are calculated by the PR curve, which are defined as follows:

$$Precision = \frac{|M \cup G|}{|M|}, \tag{16}$$

$$Recall = \frac{|M \cap G|}{|G|}, \tag{17}$$

where M is the conversion of S into a binary image, specifically, a set of thresholds are used to divide the saliency map S, and the threshold variation range is 0-255. First, the recall and precision are calculated for each threshold, then, they are combined to represent the PR curves of the model at different thresholds. The PR curves reflect the mean recall and precision of different saliency maps at various thresholds.

The F-measure generates the harmonic mean of precision P (precision) and recall R (recall), which is expressed as:

$$F_{\beta} = (1 + \beta^2)\frac{P \times R}{\beta^2 \times P + R}, \tag{18}$$

where β is set to emphasize the accuracy [48], and based on [48], F_{β}^{adp} is obtained under various thresholds (0-255).

The S-measure calculates the structural similarity between the predicted image and the ground-truth map, which is defined as:

$$S_{\alpha} = \alpha * S_o + (1 - \alpha) * S_r, \tag{19}$$

where S_o represents object-aware structural similarity, S_r represents region-aware structural similarity, and α is a balance parameter, set to 0.5.

The E-measure measures both image-level statistics and local pixel-matching information, which is defined as:

$$E_{\phi} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_{FM}(i, j), \tag{20}$$

where ϕ_{FM} represents the enhanced alignment matrix [44], W and H stand for the width and height of an image, respectively, and we can get E_{ξ}^{adp} from the basics.

MAE calculates the pixel-level errors between the ground truth G and the predicted saliency map S, which is defined as:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S_{(i,j)} - G_{(i,j)}|, \tag{21}$$

where $S(i, j)$ and $G(i, j)$ present the value of the pixel (i, j) from predicted saliency map and ground truth, respectively.

3.3. Implementation Details

We implemented our proposed methodology with one NVIDIA GeForce RTX 3090 GPU founded on the Pytorch framework. No preprocessing techniques are used to support the model during training and debugging to maximize performance. All images have been resized to 384 $\times$ 384, and three channels are copied with the depth maps. We use a pre-trained model from ImageNet [49] to initialize the parameters of the backbone network. Then, we enforce augmentation on all the training dataset to avoid overfitting, including random flipping and rotating, such results show that we can achieve better results with enhancement training in most cases. Furthermore, the learning rate initiates at a value of $1e^{-5}$, which degrades by 10 every 100 epochs. With a batch size of 3, we train our network for 150 epochs till convergence.

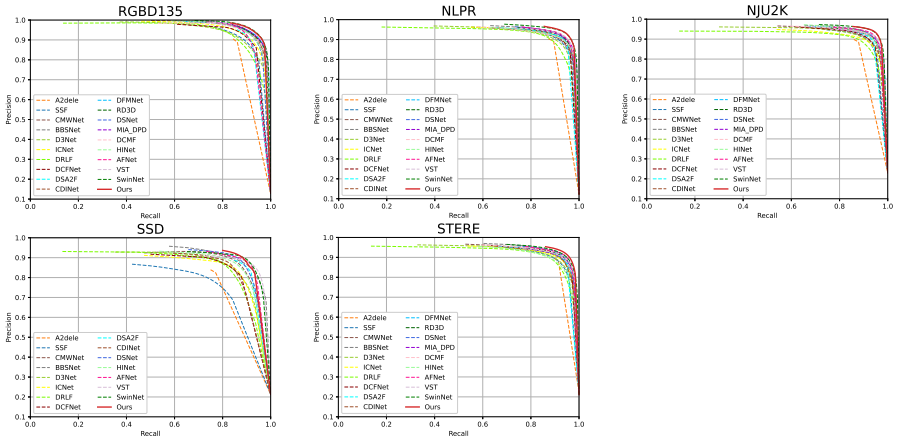

Fig. 7. PR curves of the proposed model and other salient object detection techniques on RGBD135 [37], NLPR [38], NJU2K [39], SSD [40], and STERE [41] datasets

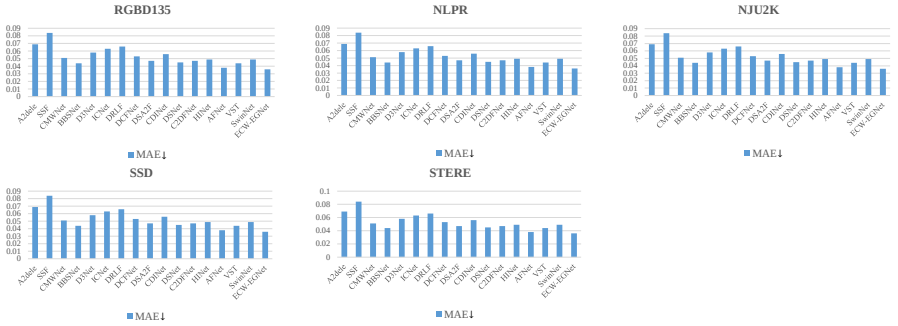

Fig. 8. MAE visualization result of the proposed model and other salient object detection techniques on RGBD135 [37], NLPR [38], NJU2K [39], SSD [40], and STERE [41] datasets

3.4. Comparison with SOTA Methods

To validate the effectiveness of the proposed ECW-EGNet, we compare with 21 SOTA methods, including A2dele [50], SSF [51], CMWNet [52], BBSNet [53], D3Net [54], ICNet [55], DRLF [6], DCFNet [56], DSA2F [57], CDINet [58], DFMNet [59], RD3D [60], DSNet [61], MIA_DPD [62], DCMF [63], C2DFNet [64], HINet [9], AFNet [7], VST [65], SwinNet [30], CATNet [66] and FCFNet [67]. For fair comparisons, the authors or providers of these methods provided or gave the default parameter settings for implementation of the original code, which was used to produce the saliency maps.

Quantitative evaluation. To validate the effectiveness of the ECW-EGNet proposed in this paper, our model is compared with 21 SOTA methods in a comprehensive manner

in terms of four evaluation metrics, with higher E_{ξ}^{adp}, S_{α} and F_{β}^{adp} values indicating better performance and the opposite for MAE values, as shown in Table 1. We can see that our method achieves the best results in almost all five test datasets. This can be attributed to three main aspects: Firstly, our CMWF module can perform a weighted fusion of low-quality depth maps. Secondly, BSCC can effectively integrate high-level semantic information and highlight object regions. Finally, EG synthesizes more accurate edge masks to guide the upgrading of salient object edges. In addition, we also provide the Precision-Recall (PR) curves in Figure 7 and visualize the MAE from Table 1 (as depicted in Figure 8) to further show the effectiveness of the proposed ECW-EGNet method. As can be seen from the results, our model achieves higher precision and recall scores on the five datasets compared to other methods. Thus, all quantitative measures demonstrate the effectiveness of the proposed model.

Table 1. Quantitative comparisons of the proposed method against the other 21 state-of-the-art RGB-D SOD methods. ↑/↓ indicates that a larger/smaller is better. The top three results are highlighted in red, blue, and green, respectively

Dataset	Metric	A2dele	SSF	CMWNet	BBSNet	D3Net	ICNet	DRLF	DCFNet	DSA2F	CDINet	DFMNet	RD3D	DSNet	MALDPD	DCMF	C^{2}DFNet	HINet	AFNet	FCFNet	VST	SwinNet	CATNet	ECW-EGNet
		CVPR20	CVPR20	ECCV20	ECCV2020	TNNLS20	TIP21	TIP21	CVPR21	CVPR21	MM21	MM21	AAAI21	TMM21	NP22	TIP22	TMM22	PR23	NP23	TSCVT23	ICCV21	TCSVT22	TMM 23	Ours
		CNN-based																			Transformer-based			
RGBD135	$E_{\xi}^{adp}\uparrow$	0.922	0.948	0.967	0.967	0.951	0.959	0.954	0.960	0.957	0.972	0.972	0.975	0.970	0.973	0.967	0.967	0.969	0.954	0.978	**0.979**	**0.980**	**0.980**	**0.977**
	$S_{\alpha}\uparrow$	0.886	0.905	0.934	0.933	0.897	0.920	0.895	0.918	0.916	0.937	0.938	0.935	0.928	0.936	0.932	0.924	0.927	0.925	**0.939**	**0.943**	**0.945**	**0.945**	**0.939**
	$F_{\beta}^{adp}\uparrow$	0.865	0.876	0.900	0.906	0.870	0.889	0.868	0.895	0.898	0.913	0.907	0.917	0.910	0.917	0.896	0.907	0.907	0.894	0.908	0.917	**0.926**	**0.923**	**0.933**
	$MAE\downarrow$	0.028	0.025	0.022	0.021	0.031	0.027	0.030	0.022	0.021	0.020	0.019	0.019	0.021	0.018	0.023	0.018	0.022	0.022	**0.017**	**0.017**	**0.016**	**0.016**	**0.015**
NLPR	$E_{\xi}^{adp}\uparrow$	0.945	0.951	0.940	0.953	0.945	0.944	0.936	0.956	0.952	0.955	0.954	0.959	0.957	0.958	0.940	0.961	0.950	0.960	0.949	0.956	**0.969**	**0.971**	**0.973**
	$S_{\alpha}\uparrow$	0.896	0.914	0.917	0.930	0.911	0.922	0.903	0.921	0.917	0.927	0.925	0.930	0.926	0.931	0.922	0.927	0.922	**0.936**	0.924	0.931	**0.941**	**0.940**	**0.941**
	$F_{\beta}^{adp}\uparrow$	0.878	0.875	0.859	0.882	0.861	0.869	0.844	0.893	0.896	0.883	0.880	0.892	0.886	0.887	0.854	0.899	0.877	0.886	0.879	0.886	**0.908**	**0.917**	**0.924**
	$MAE\downarrow$	0.028	0.026	0.029	0.023	0.029	0.028	0.032	0.023	0.024	0.024	0.024	0.022	0.024	0.022	0.029	0.021	0.026	**0.020**	0.024	0.023	**0.018**	**0.018**	**0.016**
NJU2K	$E_{\xi}^{adp}\uparrow$	0.916	0.935	0.922	0.942	0.915	0.912	0.903	0.941	0.937	0.945	0.937	0.942	0.947	0.944	0.925	0.941	0.939	0.949	0.929	0.943	**0.954**	**0.958**	**0.960**
	$S_{\alpha}\uparrow$	0.869	0.899	0.903	0.921	0.901	0.894	0.886	0.903	0.903	0.919	0.912	0.916	0.921	0.914	0.913	0.908	0.915	0.926	0.918	0.922	**0.935**	**0.932**	**0.934**
	$F_{\beta}^{adp}\uparrow$	0.874	0.886	0.880	0.902	0.865	0.867	0.849	0.898	0.901	0.907	0.894	0.901	0.907	0.898	0.881	0.899	0.896	0.909	0.891	0.900	**0.922**	**0.929**	**0.931**
	$MAE\downarrow$	0.051	0.042	0.045	0.035	0.046	0.052	0.055	0.038	0.039	0.035	0.039	0.037	0.034	0.036	0.043	0.038	0.039	0.032	0.034	0.034	**0.027**	**0.026**	**0.024**
SSD	$E_{\xi}^{adp}\uparrow$	0.870	0.873	0.902	0.920	0.904	0.879	0.880	0.906	0.912	0.906	-	-	0.923	-	-	0.920	0.916	**0.932**	-	0.922	**0.925**	-	**0.938**
	$S_{\alpha}\uparrow$	0.808	0.790	0.875	0.882	0.856	0.848	0.835	0.852	0.876	0.853	-	-	0.885	-	-	0.872	0.865	**0.897**	-	0.889	**0.892**	-	**0.891**
	$F_{\beta}^{adp}\uparrow$	0.790	0.761	0.820	0.849	0.813	0.799	0.801	0.829	0.852	0.827	-	-	0.853	-	-	0.848	0.837	**0.855**	-	0.842	**0.863**	-	**0.880**
	$MAE\downarrow$	0.069	0.084	0.051	0.044	0.058	0.063	0.066	0.053	0.047	0.056	-	-	0.045	-	-	0.047	0.049	**0.038**	-	0.044	**0.040**	-	**0.036**
STERE	$E_{\xi}^{adp}\uparrow$	0.935	0.935	0.930	0.941	0.923	0.925	0.916	0.945	0.949	0.942	0.939	0.944	0.947	0.942	0.930	0.946	0.927	**0.952**	0.927	0.942	0.950	**0.954**	**0.958**
	$S_{\alpha}\uparrow$	0.885	0.893	0.905	0.908	0.898	0.905	0.888	0.908	0.903	0.905	0.908	0.911	0.915	0.909	0.910	0.905	0.892	0.918	0.906	0.913	**0.919**	**0.921**	**0.922**
	$F_{\beta}^{adp}\uparrow$	0.884	0.880	0.869	0.885	0.859	0.864	0.845	0.897	**0.898**	0.890	0.875	0.886	0.894	0.882	0.866	0.892	0.859	**0.898**	0.868	0.878	0.893	**0.904**	**0.910**
	$MAE\downarrow$	0.043	0.044	0.043	0.041	0.046	0.045	0.050	0.037	0.036	0.041	0.040	0.037	0.036	0.037	0.043	0.038	0.049	0.034	0.034	0.038	**0.033**	**0.030**	**0.028**

Qualitative evaluation. Figure 9 shows the qualitative comparison of our proposed approach and other SOTA methods in in various challenging scenarios, including low-quality depth maps (1st and 2nd rows), small objects (3rd and 4th rows), multiple objects (5th and 6th rows), low contrast (7th and 8th rows), complex scenes (9th and 10th rows).

In the 1st and 2nd rows of the example, the depth map information is poor, but our method can still accurately detect the saliency object, which proves the effectiveness of the CMWF module. The 3rd and 4th rows show some small object samples, and the 3rd row includes a small flower. Despite the small size of these objects, our method is also able to detect them accurately. Examples with multiple objects are shown in the 5th and 6th rows. In the 6th row, most SOTA methods fail to detect three complete windows, while our method is able to segment salient objects accurately. The 7th and 8th rows demonstrate examples of low contrast between background and object regions. In the 8th row, most SOTA methods fail to separate the owl from the branch, but our method is able to accurately separate the salient object from the background. Finally, the 9th and 10th rows show examples of complex scenarios where our method accurately detects

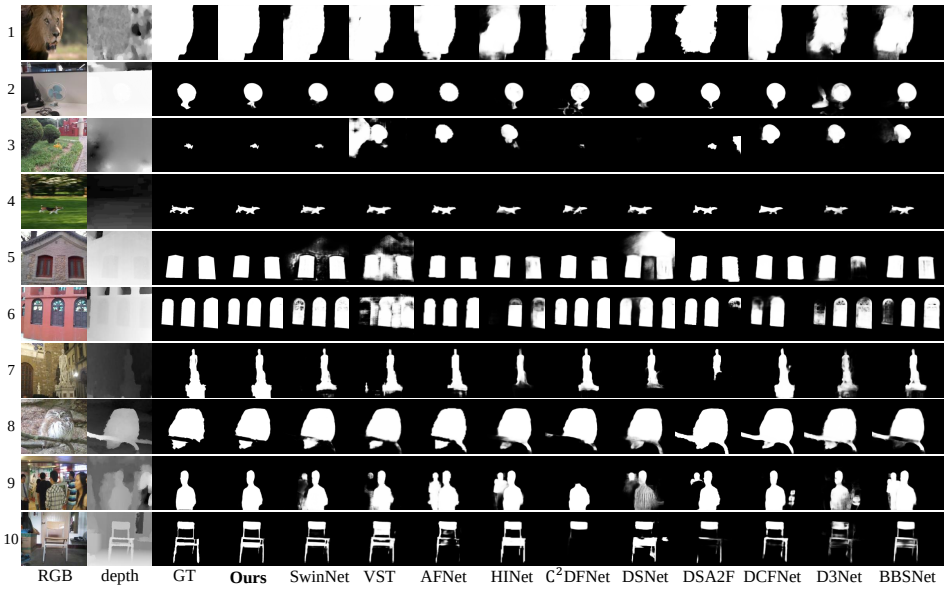

Fig. 9. Visual comparison of different RGB-D SOD methods. Including SwinNet [30], VST [65], AFNet [7], HINet [9], C2DFNet [64], DSNet [61], DSA2F [57], DCFNet [56], D3Net [54], BBSNet [53]

salient objects by effectively suppressing background interference, while other methods produce spurious results. On the 10th row, our method completely segmented the objects by highlighting foreground objects, while the other methods output incomplete results. The above examples show that our approach can effectively detect meaningful objects in a variety of scenarios.

4. Discussion

In this section, we perform ablation studies to verify the effectiveness of each component of our method on the NLPR and STERE datasets. We focus on (1) the importance of the CMWF module, (2) the effect of BSCC on RGB-D SOD, (3) the necessity of Edge-Guided (EG) decoders, and (4) the impact of multi-scale supervision (MS). As shown in Table 2, where $M1$ indicates that the model does not use any components and only contains MS. $M2$ added CMWF and MS, and removed BSCC and EG. $M3$ means that BSCC and MS are added to the model, and CMWF and EG are removed. $M4$ indicates that the model contains both CMWF, EG, and MS, but does not use BSCC. $M5$ indicates that the model contains both CMWF, BSCC, and EG, but does not use MS. $M6$ means that the model uses all components simultaneously (i.e., CMWF, BSCC, EG, and MS).

The effectiveness of CMWF. The CMWF module merges RGB and depth images in our network in a weighted manner. By comparing $M1$ and $M2$, we can see a significant improvement in performance for both datasets. This phenomenon is even more pronounced in the STERE dataset: the F-measure and MAE of the CMWF module im-

Table 2. Ablation studies for our ECW-EGNet on two datasets. " $\surd$ " denote the model contains the corresponding component

Model	CMWF	BSCC	EG	MS	NLPR				STERE			
					$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{adp}\uparrow$	$MAE\downarrow$	$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{adp}\uparrow$	$MAE\downarrow$
M 1				$\surd$	0.965	0.936	0.907	0.020	0.942	0.917	0.887	0.035
M 2	$\surd$			$\surd$	0.969	0.937	0.917	0.018	0.952	0.920	0.902	0.031
M 3		$\surd$		$\surd$	0.972	0.940	0.918	0.017	0.953	0.921	0.902	0.030
M 4	$\surd$		$\surd$	$\surd$	0.970	0.941	0.917	0.017	0.951	0.922	0.902	0.030
M 5	$\surd$	$\surd$	$\surd$		0.969	0.941	0.921	0.017	0.955	0.921	0.904	0.030
M 6	$\surd$	$\surd$	$\surd$	$\surd$	**0.973**	**0.941**	**0.924**	**0.016**	**0.958**	**0.922**	**0.910**	**0.028**

prove performance by 1.7% and 11.4% respectively. This is because STERE contains a high volume of noisy depth images, whereas our CMWF reduces the interference information from low-quality images. To demonstrate the superiority of CMWF, we have experimentally compared the performance of three variants of the CMWF module. Specifically, the model $w/o\ CMWF$ express our full model without the CMWF module; The model $w/o\ CSAE$ illustrate our full model without the CSAE mechanism; and the model $w/o\ DQA$ displays our full model without the DQA mechanism. The results are shown in Table 3 and it can be seen that all three steps in CMWF are helpful in improving the performance of SOD, especially after removing the DQA module there is a clear decrease in performance, which confirms the potential of our DQA to evaluate depth images and reduce the negative impact of low-quality depth maps.

The effectiveness of BSCC. In ECW-EGNet, multi-scale feature fusion is achieved by using depth-wise separable convolutions of various convolutional kernel sizes. In Table 2, the comparison between $M1$ and $M3$ shows that the addition of the BSCC module to integrate global features improves the performance of the network on these two datasets. Specifically, when evaluating the NLPR dataset, the E-measure and F-measure of the model increase by 0.7% and 1.2% respectively upon incorporating the BSCC module.

Table 3. Comparison of quantitative indicators between CMWF and variant

Model	NLPR				STERE			
	$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{adp}\uparrow$	$MAE\downarrow$	$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{adp}\uparrow$	$MAE\downarrow$
w/o CMWF	0.962	0.930	0.904	0.020	0.948	0.918	0.895	0.033
w/o CSAE	0.972	0.940	0.918	0.017	0.953	0.918	0.901	0.031
w/o DQA	0.971	0.940	0.921	0.017	0.951	0.921	0.901	0.031
CMWF	**0.973**	**0.941**	**0.924**	**0.016**	**0.958**	**0.922**	**0.910**	**0.028**

The effectiveness of EG. It refines the contour details of salient objects by extracting salient edge features and integrating the edge features into the decoding process. In Table 2, by looking in $M2$ and $M4$, we can see that adding EG also wins edge segmentation on NLPR, e.g., E_ξ^{adp}: 0.969$\longrightarrow$ 0.970, S_α: 0.937$\longrightarrow$ 0.941, MAE: 0.018$\longrightarrow$ 0.017.

Effectiveness of multi-scale supervision (MS). During the training of the network, we uses multi-scale supervision to achieve an accurate representation of the salient objects. To verify the effectiveness of multiple supervision, we remove all other losses in the network, keeping only the last layer of loss L1. The results for this experiment are presented in row $M5$ of Table 2. It can be observed that the removal of multi-scale supervision results in a decrease in performance, which confirms that the introduction of supervision in the output of each side facilitates performance improvement.

5. Conclusions

In this paper, we propose an efficient ECW-EGNet framework to implement RGB-D SOD. We propose a cross-modality weighting fusion (CMWF) module that assesses the quality of the depth map and fuses cross-modal features by calculating the difference between RGB modal and depth modal in order to compensate for the difference between RGB and depth modes. In CMWF, features extracted from the backbone network are enhanced to generate discriminative features, which are then used to generate weighting factors for the enhanced discriminative features to determine the weights for depth feature fusion. Additionally, we introduce a bi-directional scale-correlation convolution (BSCC) module to learn high-level semantic information to better capture contextual information for effective guided saliency prediction. Moreover, the accuracy of SOD is further improved with the use of an Edge-Guided (EG) decoder. Experimental tests on five representative datasets show that the proposed method outperforms 21 SOTA methods in four evaluation metrics. In the near future, we are especially interested in the prospect of reducing model parameters without compromising model performance. We design a more lightweight backbone network to replace the current Swin Transformer for feature extraction.

Acknowledgments. This work was supported by the Anhui University of Science and Technology Medical Special Training Project(YZ2023H2B003), Huainan City Science and Technology Plan Project(2023A316), National Natural Science Foundation of China (62102003), Anhui Postdoctoral Science Foundation (2022B623), Natural Science Foundation of Anhui Province (2108085QF258), the University Synergy Innovation Program of Anhui Province (GXXT-2021-006, GXXT-2022-038), Academic Support Program for Top Talents in Academic Disciplines (Majors) of Anhui Province (gxbjZD2021050), the Institute of Energy, Hefei Comprehensive National Science Center under (21KZS217), University-level general projects of Anhui University of science and technology (xjyb2020-04), Central guiding local technology development special funds (202107d06020001).

References

1. Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7223–7233, 2019.
2. Wei Xiong, Yafei Lv, Yaqi Cui, Xiaohan Zhang, and Xiangqi Gu. A discriminative feature learning approach for remote sensing image retrieval. *Remote Sensing*, 11(3):281, 2019.
3. Junyang Yu, Mengle Zuo, Lifeng Dong, Huanlong Zhang, and Xin He. The multi-level classification and regression network for visual tracking via residual channel attention. *Digital Signal Processing*, 120:103269, 2022.

4. Fenglei Chen, Haijun Liu, Zhihong Zeng, Xichuan Zhou, and Xiaoheng Tan. Bes-net: Boundary enhancing semantic context network for high-resolution image semantic segmentation. *Remote Sensing*, 14(7):1638, 2022.
5. Kang Yi, Jinchao Zhu, Fu Guo, and Jing Xu. Cross-stage multi-scale interaction network for rgb-d salient object detection. *IEEE Signal Processing Letters*, 29:2402–2406, 2022.
6. Xuehao Wang, Shuai Li, Chenglizhao Chen, Yuming Fang, Aimin Hao, and Hong Qin. Data-level recombination and lightweight fusion scheme for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:458–471, 2020.
7. Tianyou Chen, Jin Xiao, Xiaoguang Hu, Guofeng Zhang, and Shaojie Wang. Adaptive fusion network for rgb-d salient object detection. *Neurocomputing*, 522:152–164, 2023.
8. Chenxing Xia, Songsong Duan, Xianjin Fang, Xiuju Gao, Yanguang Sun, Bin Ge, Hanling Zhang, and Kuan-Ching Li. Efgnet: Encoder steered multi-modality feature guidance network for rgb-d salient object detection. *Digital Signal Processing*, 131:103775, 2022.
9. Hongbo Bi, Ranwan Wu, Ziqi Liu, Huihui Zhu, Cong Zhang, and Tian-Zhu Xiang. Cross-modal hierarchical interaction network for rgb-d salient object detection. *Pattern Recognition*, 136:109194, 2023.
10. Zhihong Zeng, Haijun Liu, Fenglei Chen, and Xiaoheng Tan. Compensated attention feature fusion and hierarchical multiplication decoder network for rgb-d salient object detection. *Remote Sensing*, 15(9):2393, 2023.
11. Chenglizhao Chen, Jipeng Wei, Chong Peng, and Hong Qin. Depth-quality-aware salient object detection. *IEEE Transactions on Image Processing*, 30:2350–2363, 2021.
12. Hao Chen, Yongjian Deng, Youfu Li, Tzu-Yi Hung, and Guosheng Lin. Rgbd salient object detection via disentangled cross-modal fusion. *IEEE Transactions on Image Processing*, 29:8407–8416, 2020.
13. Xiaolong Cheng, Xuan Zheng, Jialun Pei, He Tang, Zehua Lyu, and Chuanbo Chen. Depth-induced gap-reducing network for rgb-d salient object detection: an interaction, guidance and refinement approach. *IEEE Transactions on Multimedia*, 2022.
14. Chenxing Xia, Songsong Duan, Xianjin Fang, Bin Ge, Xiuju Gao, and Jianhua Cui. Dast: Depth-aware assessment and synthesis transformer for rgb-d salient object detection. In *Pacific Rim International Conference on Artificial Intelligence*, pages 473–487, 2022.
15. Qian Chen, Keren Fu, Ze Liu, Geng Chen, Hongwei Du, Bensheng Qiu, and Ling Shao. Ef-net: A novel enhancement and fusion network for rgb-d saliency detection. *Pattern Recognition*, 112:107740, 2021.
16. Gongyang Li, Zhi Liu, Minyu Chen, Zhen Bai, Weisi Lin, and Haibin Ling. Hierarchical alternate interaction network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:3528–3542, 2021.
17. Yang Yang, Qi Qin, Yongjiang Luo, Yi Liu, Qiang Zhang, and Jungong Han. Bi-directional progressive guidance network for rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5346–5360, 2022.
18. Chenglizhao Chen, Jipeng Wei, Chong Peng, Weizhong Zhang, and Hong Qin. Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion. *IEEE Transactions on Image Processing*, 29:4296–4307, 2020.
19. Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. Cdnet: Complementary depth network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:3376–3390, 2021.
20. Wujie Zhou, Qinling Guo, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Irfr-net: Interactive recursive feature-reshaping network for detecting salient objects in rgb-d images. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
21. Yu Qiu, Yun Liu, Yanan Chen, Jianwen Zhang, Jinchao Zhu, and Jing Xu. A2sppnet: attentive atrous spatial pyramid pooling network for salient object detection. *IEEE Transactions on Multimedia*, 2022.

22. Nianchang Huang, Yi Liu, Qiang Zhang, and Jungong Han. Joint cross-modal and unimodal features for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 23:2428–2441, 2020.
23. Wenbo Zhang, Yao Jiang, Keren Fu, and Qijun Zhao. Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2021.
24. Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Rrecognition*, pages 3052–3062, 2020.
25. Zhengyi Liu, Kaixun Wang, Hao Dong, and Yuan Wang. A cross-modal edge-guided salient object detection for rgb-d image. *Neurocomputing*, 454:168–177, 2021.
26. Zhengyi Liu, Song Shi, Quntao Duan, Wei Zhang, and Peng Zhao. Salient object detection for rgb-d image by single stream recurrent convolution neural network. *Neurocomputing*, 363:46–57, 2019.
27. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, 2021.
28. Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, 23(2):358–367, 1988.
29. Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In *Proceedings of the International Conference on Multimedia*, pages 4481–4490, 2021.
30. Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4486–4497, 2021.
31. Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13756–13765, 2020.
32. Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *Proceeding of the European Conference on Computer Vision*, pages 52–69. Springer, 2020.
33. Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018.
34. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.
35. Yu-Huan Wu, Yun Liu, Le Zhang, Ming-Ming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, 31:3125–3136, 2022.
36. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceeding of the Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.
37. Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, pages 23–27, 2014.
38. Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *Proceeding of the European Conference on Computer Vision*, pages 92–109, 2014.

39. Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *IEEE International Conference on Image Processing*, pages 1115–1119, 2014.
40. Chunbiao Zhu and Ge Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3008–3014, 2017.
41. Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 454–461, 2012.
42. Yingjie Zhai, Deng-Ping Fan, Jufeng Yang, Ali Borji, Ling Shao, Junwei Han, and Liang Wang. Bifurcated backbone strategy for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:8727–8742, 2021.
43. Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4681–4691, 2021.
44. Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018.
45. Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.
46. Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4548–4557, 2017.
47. Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–740, 2012.
48. Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics*, 48(11):3171–3183, 2017.
49. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
50. Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan 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*, pages 9060–9069, 2020.
51. Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for rgb-d saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3472–3481, 2020.
52. Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for rgb-d salient object detection. In *Proceeding of the European Conference on Computer Vision*, pages 665–681, 2020.
53. Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *Proceeding of the European Conference on Computer Vision*, pages 275–292, 2020.
54. Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2075–2089, 2020.
55. Gongyang Li, Zhi Liu, and Haibin Ling. Icnet: Information conversion network for rgb-d based salient object detection. *IEEE Transactions on Image Processing*, 29:4873–4884, 2020.

56. Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9471–9481, 2021.
57. Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1407–1417, 2021.
58. Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam Kwong. Cross-modality discrepant interaction network for rgb-d salient object detection. In *Proceedings of the International Conference on Multimedia*, pages 2094–2102, 2021.
59. Wenbo Zhang, Ge-Peng Ji, Zhuo Wang, Keren Fu, and Qijun Zhao. Depth quality-inspired feature manipulation for efficient rgb-d salient object detection. In *Proceedings of the International Conference on Multimedia*, pages 731–740, 2021.
60. Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. Rgb-d salient object detection via 3d convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 2, pages 1063–1071, 2021.
61. Hongfa Wen, Chenggang Yan, Xiaofei Zhou, Runmin Cong, Yaoqi Sun, Bolun Zheng, Jiyong Zhang, Yongjun Bao, and Guiguang Ding. Dynamic selective network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:9179–9192, 2021.
62. Yanhua Liang, Guihe Qin, Minghui Sun, Jun Qin, Jie Yan, and Zhonghan Zhang. Multi-modal interactive attention and dual progressive decoding network for rgb-d/t salient object detection. *Neurocomputing*, 490:132–145, 2022.
63. Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:1285–1297, 2022.
64. Miao Zhang, Shunyu Yao, Beiqi Hu, Yongri Piao, and Wei Ji. C^2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 2022.
65. Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4722–4732, 2021.
66. Fuming Sun, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, pages 1–14, 2023.
67. Qiang Zhang, Qi Qin, Yang Yang, Qiang Jiao, and Jungong Han. Feature calibrating and fusing network for rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.

Chenxing Xia received the Ph.D. degree from Hunan University in 2019. He is currently an Associate Professor at the College of Computer Science and Engineering, Anhui University of Science and Technology. His research interests include saliency detection, computer vision, image processing, and depth prediction.

Feng Yang received the bachelor degree from Anhui Sanlian University. She is currently pursuing the master degree with Anhui University of Science and Technology. Her research interests include computer vision, machine learning, salient objection detection.

Songsong Duan received the M.S. degree with Anhui University of Science and Technology. He is currently pursuing the Ph.D. degree with the School of Telecommunications Engineering, Xidian University, China. His research interests include computer vision, salient object detection, and weakly supervised learning.

Xiuju Gao received the M.S. degree in Computer Science and Electronic Engineering from Hunan University, Changsha China, in 2016. She is currently an assistant at the College of Electrical and Information Engineering, Anhui University of Science and Technology. Her research interests include image/video processing and computer vision.

Bin Ge received the Ph.D. degree in computer application technology from Hefei university of technology, Hefei China, in 2016. He began to teach at Anhui University of Science and Technology in 1999. Currently, he is a professor of information security at Anhui University of Science and Technology. His research interests include information security, Internet of things technology and artificial intelligence.

Kuan-Ching Li is a Distinguished Professor in the Dept of Computer Science and Information Engineering (CSIE) at Providence University, Taiwan, where he also serves as the Director of the High-Performance Computing and Networking Center. He received the Licenciatura in Mathematics, and MS and Ph.D. degrees in electrical engineering from the University of Sao Paulo (USP), Brazil, in 1994, 1996, and 2001, respectively. He published more than 250 scientific papers and articles and is author, co-author or editor of more than 25 books published by Taylor & Francis, Springer, and McGraw-Hill. Professor Li is the Editor in Chief of the Connection Science and serves as an associate editor for several leading journals. Also, he has been actively involved in many major conferences and workshops in program/general/steering conference chairman positions and has organized numerous conferences related to computational science and engineering. He is a Fellow of IET and a Senior Member of the IEEE. His research interests include parallel and distributed computing, Big Data, and emerging technologies.

Xianjin Fang received the Ph.D. degree in computer application technology from Anhui University in 2010. He is a professor and Ph.D. supervisor at Anhui University of Science and Technology. His research interests include information security, data mining, etc.

Yan Zhang received the Ph.D. degree in computer application technology from Anhui University in 2010. She is a professor and Ph.D. supervisor at Anhui University of Science and Technology. Her research interests include information security, data mining, etc.

Ke Yang graduated from Huainan University of Technology with a bachelor's degree in 2001, and received the M.S. degree and Ph.D. degree from Anhui University of Science and Technology in 2004 and 2007. He is currently the dean of the School of Energy and Safety at Anhui University of Science and Technology, a second-level professor, and a doctoral supervisor. His research interests include data Analysis, mine pressure and control, special mining methods, and coal and gas co-mining.

Received: December 06, 2023; Accepted: March 01, 2024.