IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 13, 2020, accepted November 4, 2020, date of publication November 9, 2020, date of current version November 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036681

Cross-Modal Feature Integration Network for Human Eye-Fixation Prediction in RGB-D Images

WENYU LIU[1], WUJIE ZHOU[1,2], (Member, IEEE), AND TING LUO[3]

[1]School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China
[2]Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China
[3]College of Science and Technology, Ningbo University, Ningbo 315211, China

Corresponding author: Wujie Zhou (wujiezhou@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502429 and Grant 61972357, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY18F020012.

ABSTRACT With the advent of convolutional neural networks, research progress in visual saliency prediction has been impressive. While integrating features at different stages from the backbone network is important, feature extraction itself is equally relevant. A network may lose representative information during feature extraction. We address the loss of spatial information and perform a fusion of features extracted from RGB and depth data for eye-fixation prediction. Specifically, we propose an asymmetric feature extraction network comprising an edge guidance module (EGM) and a feature integration module (FIM) that processes RGB-D images. Edge guidance supports the extraction of spatial information, while feature integration merges features from RGB images and the corresponding depth maps. We obtain the eye-fixation prediction maps by linearly fusing the features from the backbone network with those optimized using the two modules. Experimental results on NCTU and NUS, two benchmark datasets for RGB-D saliency prediction, verify the effectiveness and high-performance of the proposed network compared with similar methods.

INDEX TERMS Convolutional neural network (CNN), human eye-fixation prediction, asymmetric feature extraction, edge guidance module (EGM), feature integration module (FIM).

I. INTRODUCTION

The attention mechanism in the visual system enables humans to filter out redundant or irrelevant information from massive visual data and automatically select the most outstanding or interesting areas in a visual scene for further processing. Without the visual attention mechanism, we would not be able to process the rich information from scenes we see every day. The ability to process information mimicking the visual attention mechanism is known as saliency detection in computer vision. As a preprocessing step, saliency detection promotes efficiency in a wide variety of vision-oriented multimedia applications, such as semantic segmentation [1]–[5], image quality assessment [6], [7], image and video compression [8], [9], image retargeting [10], [11], image classification [12]–[14], and image retrieval [15], [16]. Saliency detection can be divided into eye-fixation prediction [17]–[20], which predicts the focus of the human gaze, and salient object detection [21]–[25], which extracts the most salient objects or regions from a scene. In this study, we focused on eye fixation.

The associate editor coordinating the review of this manuscript and approving it for publication was Yunjie Yang.

Over the past two decades, many methods for eye-fixation prediction have been proposed. These methods can either be based on learning using handcrafted features and low-level cues or adopt deep learning to leverage convolutional neural networks (CNNs) and high-level semantic features. The traditional methods often fail to achieve high-performance image processing by the difficulty of achieving effective feature extraction. For instance, Harel *et al.* [26] proposed a simple model called Graph-Based Visual Saliency (GBVS) that forms activation maps on certain feature channels and then highlights conspicuous areas and admits combination with other maps for normalization. Zhang *et al.* [27] proposed the Saliency detection by combining simple priors (SDSP) network that combines three simple priors: frequency priors, color priors, and location priors. Qi *et al.* [28] proposed a model based on bandpass filtering to resemble visual perceptual processing. On the other hand, CNNs have broken the deadlock given their excellent feature extraction ability that allows independent extraction of effective features for different images and extracts high-level semantic information from images. Simonyan and Zisserman [29] investigated the effect of the CNN depth on accuracy in a large-scale image recognition setting and proposed a

very deep CNN to this end. He *et al.* [30] pointed out that deeper networks are more difficult to train, and thus developed a residual learning framework to ease training. Huang *et al.* [31] introduced a dense convolutional network that links every pair of layers with a feedforward connection.

In recent years, several models have been proposed to solve various problems in eye-fixation prediction. Lang *et al.* [32] and Yao and Hang [33] addressed the influence of depth cues on visual saliency. Cornia *et al.* [34] proposed a deep multilevel network for saliency prediction. Jia and Bruce [35] introduced an expandable multilayer network for saliency prediction. Kruthiventi *et al.* [36] devised a fully CNN for predicting human eye fixation. Yang *et al.* [37] proposed a dilated inception network for visual saliency prediction. Lv *et al.* [38] developed an attention-based fusion network for human eye-fixation prediction in 3-D images. Cheng *et al.* [39] proposed a computational model for stereoscopic visual saliency prediction. Some of these methods are based on RGB single-stream input data, and others are based on RGB-D dual-stream input data. Considering that depth maps contain abundant spatial and shape information, they complement and are highly correlated with RGB images. Therefore, compared with RGB input data, RGB-D data are being increasingly adopted in computer vision research [28], [40]–[43].

We propose a CNN based on RGB-D data for eye-fixation prediction. Directly fusing RGB and depth information is usually insufficient to complete challenging computer vision tasks. Zhang *et al.* [44] proposed a linear fusion strategy to integrate RGB images and depth maps. Likewise, Wang *et al.* [45] proposed a visual attention module for their fusion. Jiang *et al.* [46] proposed an attention mechanism to allocate different weights to multilevel RGB and depth features and obtain fusion features. Moreover, extracting feature information from a basic network may not be effective enough. To address this problem, Li *et al.* [47] manually produced features from four modes to compensate for the disadvantages of a CNN. We devised an asymmetric feature extraction framework to extract representative information from RGB images and depth maps. As RGB images contain more information than depth maps, we use different backbone frames to extract complementary information from these two types of data. Specifically, we use VGGNet16 as the encoding network for RGB images and ResNet34 as the encoding network for depth maps. Then, two feature optimization blocks, the edge guidance module (EGM) and the feature integration module (FIM), are introduced. We consider the features from the first two layers as low-level features and those from the last three layers as high-level features. Low-level features in shallow layers can retain spatial information to reconstruct object boundaries, while high-level features in the deep layers encode semantic information to obtain object descriptions. The EGM complements spatial information from the backbone network, and the FIM enhances the feature fusion between RGB and depth data.

Our contributions can be summarized as follows:

(1) Instead of using the symmetric network, we propose an asymmetric network that processes RGB images and depth maps to extract the features using the VGGNet16 and ResNet34 networks, respectively, as a backbone framework for feature extraction.

(2) The EGM ensures that features from the RGB image and depth maps are fully extracted and then combined using the FIM. By combining a spatial attention mechanism and cross-modal fusion, the FIM provides a feature map with detailed characteristics.

(3) The proposed network outperforms state-of-the-art methods on two benchmark datasets for eye-fixation prediction.

II. PROPOSED NETWORK FOR HUMAN EYE-FIXATION PREDICTION

The proposed network architecture includes four parts, namely, feature extraction backbone network, FIM, EGM, and decoder, as shown in Figure 1. The feature extraction backbone is modified from the VGGNet16 [29] and ResNet34 [30] networks. The FIM refines and enhances the fusion of features from RGB images and depth maps. The EGM restores spatial information lost during feature extraction. Finally, the decoder combines the information from each part and gradually recovers the resolution of the original image to generate the final prediction map corresponding to human eye fixation. The four parts are detailed in Sections II.A–II.D. Then, we introduce the loss function for network training in Section II.E.

A. BACKBONE NETWORK

VGGNet and ResNet are among the most commonly used pre-trained backbone networks for human eye-fixation prediction by their unique advantages. Both networks have their extended versions. VGGNet networks include VGGNet11, VGGNet13, VGGNet16, and VGGNet19. We select VGGNet16 as the backbone network for RGB feature extraction. Similarly, ResNet networks include ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152. We use ResNet34 instead of the more common ResNet50 as the backbone network for depth feature extraction. As ResNet50 is deeper than ResNet34, we consider sufficient to extract depth information using the latter. In fact, for a learning task, once a network provides accurate results, further deepening the network provides no clear learning improvement. Instead, increasing the number of parameters by adding more layers to a network notably increases the computational burden.

As the VGGNet and ResNet networks were originally intended for image classification, we adapt their structure to perform human eye-fixation prediction. For VGGNet16, we retain the first five convolutional layers (conv1_2, conv2_2, conv3_3, conv4_3, and conv5_3) and remove the last two fully connected layers. An RGB input image of size $M \times N$ provides a feature map of size $\frac{W}{2^5} \times \frac{H}{2^5}$ after passing through the five convolutional layers. For ResNet34,

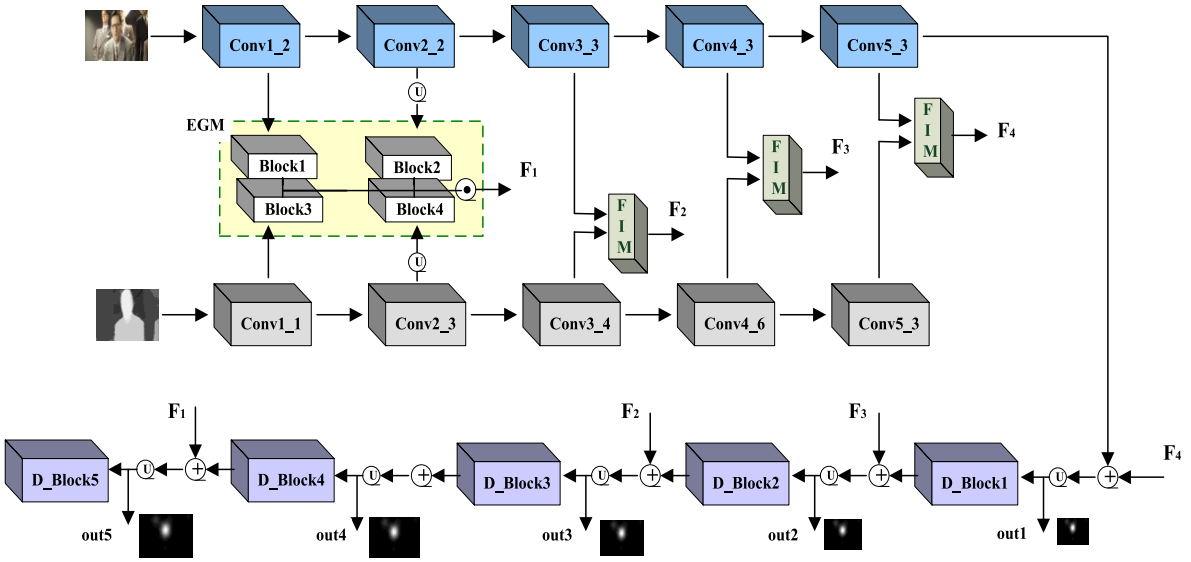

FIGURE 1. Overview of proposed network.

we retain the residual block of the convolutional layer and remove the average pooling and the last fully connected layer. The first 7×7 convolution in ResNet34 constitutes the first layer, and the following pooling layer and three residual blocks correspond to the second layer. These blocks are denoted as Conv1_1, Conv2_3, Conv3_4, Conv4_6, and Conv5_3, respectively. Similarly, a depth map of size $M \times N$ provides a feature map of size $\frac{M}{2^5} \times \frac{N}{2^5}$.

B. EGM

Spatial information is lost as the number of pooling operations increases. Although some methods use attention mechanisms to prevent this problem [48]–[52], we consider that the loss of spatial characteristics cannot be completely avoided when a network implements more than four pooling operations. Zhang and Pang [53] noted that edge information can provide useful fine-grained constraints to guide feature extraction during segmentation, and only low-level features can retain detailed edge information. Edge information not only plays an important role in segmentation, but it is also indispensable for human eye-fixation prediction. Edge features provide information on the position, size, and shape of salient objects. Therefore, we implement the proposed EGM at the top of the first two layers to learn an edge attention representation during early encoding and retain local edge features.

EGM architecture is shown in Figure 2. Block1–Block4 are called edge guidance blocks and comprise 1×1 and 3×3 convolutions. The outputs from layers conv2_2 and conv2_3 are first upsampled to the same resolution of layers conv1_2 and conv1_1 and then fed into the edge guidance blocks. The feature after splicing, F_1, is given by

$$F_{rgb} = f_{3\times3}(f_{1\times1}(f_{r_1}))) \cdot f_{3\times3}(f_{1\times1}(U(f_{r_2})))), \quad (1)$$

$$F_{depth} = f_{3\times3}(f_{1\times1}(f_{d_1}))) \cdot f_{3\times3}(f_{1\times1}(U(f_{d_2})))), \quad (2)$$

$$F_1 = F_{rgb} \cdot F_{depth}, \quad (3)$$

where $\cdot$ represents concatenation, U represents upsampling with a scale factor of 2, f_{r_i} and f_{d_i} $(i = 1, 2, 3)$ denote the feature maps from the first two layers in the RGB and depth networks, respectively, and $f_{1\times1}$ and $f_{3\times3}$ denote convolutions with kernels 1×1 and 3×3, respectively. During decoding, F_1 is fused with the features in the decoding block to enhance both spatial information and the corresponding human eye-fixation prediction.

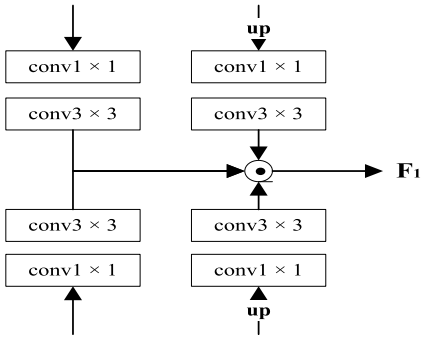

FIGURE 2. Architecture of proposed EGM.

C. FIM

RGB images convey visual characteristics by color discrimination, while depth maps represent the geometry of objects. Thus, properly integrating features from RGB and depth data is essential for human eye-fixation prediction given their strong correlation and complementarity. However, depth information also contains noise and interference, which may be exacerbated by inappropriate fusion. Chen and Li [54] proposed a block for the fusion of complementary perception features. Huang *et al.* [55] introduced a deep CNN to gradually fuse features from high- and low-level layers. Chen *et al.* [56] performed a fusion of multiscale, multipath, and multimode features. Likewise, the proposed FIM enables the fusion of features from RGB images and depth maps.

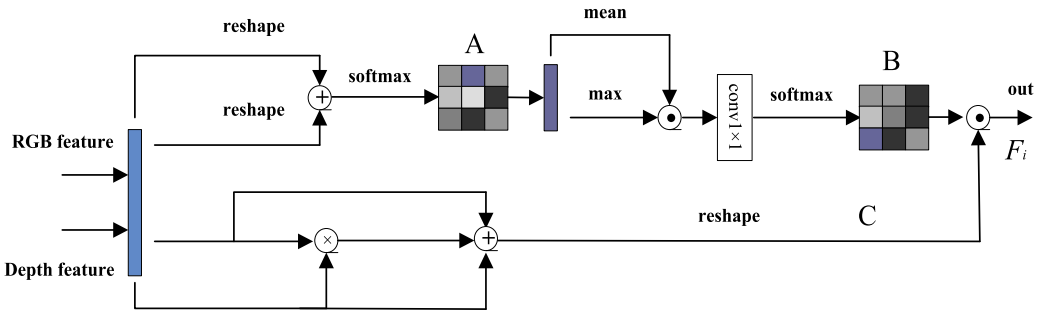

FIGURE 3. Architecture of proposed FIM.

The detailed FIM architecture is shown in Figure 3. The FIM consists of an attention mechanism for mining detailed spatial information and cross-modal fusion for improving the scope of the observable information in the feature map. Let sets A and B denote the features from the attention mechanism and cross-modal fusion, respectively. They are computed as

$$A = soft\ \max(reshape(f_{rgb}) \oplus reshape(f_{depth})), \quad (4)$$

$$B = soft\ \max(f_{1\times1}(Mean(A) \cdot Max(A))), \quad (5)$$

$$C = reshape(f_{rgb} \otimes f_{depth}) \cdot (f_{rgb} \oplus f_{depth}), \quad (6)$$

$$F_i = B \cdot C, \quad (7)$$

where F_i ($i = 2, 3, 4$) represents the feature maps from the last three layers in the backbone networks of RGB and depth data, $\otimes$ denotes multiplication, $\oplus$ denotes addition, *Mean* and *Max* represent the arithmetic mean and maximum value of their arguments, respectively, f_{rgb} represents the features extracted from the RGB image, and f_{depth} represents the features extracted from the depth map. A new feature map is thus obtained by fusing the features obtained from the attention mechanism and cross-modal fusion. This map contains improved salient object features, spatial information, and the final map reflecting human eye-fixation prediction.

D. DECODING NETWORK

The encoding process, provides salient object features from RGB-D data, and the EGM and FIM refine and complement the extracted features. Generating the final prediction map requires a decoder network to handle the obtained features. Decoder networks are mostly constructed by superposition and deconvolution. Although deconvolution helps to map rough feature maps onto dense feature maps, it is computation-intensive and may introduce uneven patterns in some non-dense feature maps. As the proposed network considers the importance of spatial information, we use bilinear interpolation upsampling for the decoder network.

As shown in Figure 1, we implement five decoding blocks, D-Block$_i$ ($i = 1, \ldots, 5$). Each decoding block performs bilinear interpolation with a scale factor of 2, a 1×1 convolution, and a 3×3 convolution. Bilinear interpolation gradually restores the image resolution, and the convolutions integrate the features. The features from each mode (i.e., RGB and depth modes) are gradually fused during each decoding phase, and the edge information obtained from the EGM is integrated during the final decoding phase.

E. LOSS FUNCTION

We use the correlation coefficient (CC) and mean squared error in the loss function between the ground truth and eye-fixation prediction maps. The CC ranges from -1 to 1 and reflects the linear relation between two distributions. As shown in Figure 1, the proposed network provides five outputs with different scales. We calculate the ground truth of each output using the loss function. Then, we obtain the final loss by averaging the values obtained from the five loss functions as follows:

$$LOSS = \frac{1}{5}\left(\sum_{k=1}^{5}\left(\frac{1}{M}\sum_{i=1}^{m}||P-G||^2 + 1 - \frac{\sigma(P,G)}{\sigma(P)\times\sigma(G)}\right)\right), \quad (8)$$

where $\sigma(\cdot)$ represent the standard deviation of the input, $||P-G||^2$ denotes the square error, M is the number of pixels per image, k denotes the number of loss function, which corresponds to the number of outputs we have, i indexes the i^{th} pixel, index m ranges from 1 to M, $\sigma(P, G)$ denotes the covariance of P and G, and P and G represent the prediction and ground-truth maps, respectively.

III. EXPERIMENTS AND RESULTS

In this section, we report the implementation details of the proposed network, datasets and indicators for evaluation, ablation study, and experimental results.

A. IMPLEMENTATION DETAILS

We implemented and trained the proposed network on the PyTorch 1.1.0 environment running on a computer equipped with a NVIDIA TITAN Xp GPU with 12 GB of memory. We used the first five layers of the pre-trained VGGNet16 as the encoder network for RGB images, and the first five layers of the pre-trained ResNet34 as the encoder network for depth maps. The initial image input size of the model was 224×224. The proposed network was trained in an end-to-end scheme with a total of five outputs with corresponding image sizes of 14×14, 28×28, 56×56, 112×112, and 224×224. The final output was used as the final saliency prediction map, and the four intermediate

TABLE 1. Performance of proposed network with and without EGM and FIM. *-EGM* represents the removal of EGM. *-FIM* represents the removal of FIM. *-EGM & FIM* represents the removal of EGM and FIM. *-pyramid supervision* represents the model without pyramid supervision. *-All module* represents the model without pyramid supervision, and the removal of EGM and FIM.

Datasets	Criteria	*-EGM*	*-FIM*	*-EGM & FIM*	*-Pyramid supervision*	*-All module*	Proposed
NCTU	CC	0.8179	0.8165	0.8099	0.8280	0.7947	**0.8321**
	KLDiv	0.4079	0.3298	0.3149	0.2974	0.3290	**0.2832**
	AUC	0.8735	**0.8781**	0.8748	0.8804	0.8690	0.8777
	NSS	1.9113	1.8998	1.8861	1.9269	1.8451	**1.9433**
NUS	CC	0.5369	0.5289	0.5356	0.5352	0.5283	**0.5498**
	KLDiv	1.4061	1.2791	**1.1763**	1.2208	1.1808	1.2251
	AUC	0.8507	**0.8650**	0.8611	0.8384	0.8385	0.8532
	NSS	2.2297	2.0974	2.1377	2.2183	2.1462	**2.2877**

outputs were considered during the training of the multiscale supervised network to improve the prediction performance. To speed up convergence and improve efficiency for training, the regions in each image were normalized with the mean value from the RGB channel being the center before the initial weight was input into the network. The initial learning rate was 1×10^{-4}, and the batch size was one. When our network performed well in the training set but poorly in the verification set, we terminated training in advance, because this behavior indicates overfitting. In this case, if training proceeds, the network performance cannot be improved, and the generalization ability may be undermined.

B. DATASETS

To effectively evaluate the performance of the proposed network, we conducted tests on NUS and NCTU, two benchmark datasets for human eye-fixation prediction. The NCTU and NUS datasets, provided by The National Jiaotong University and the National University of Singapore, contain 475 and 600 images captured from different scenes, respectively. These datasets have been extensively used to evaluate human eye-fixation prediction. Likewise, we used these datasets for comparison of the proposed network with other methods. From the 600 images in the NUS dataset, 420 were used for training, 60 for verification, and 120 for testing. From the 475 images in the NCTU dataset, 332 were used for training, 48 for verification, and 95 for testing.

C. EVALUATION MEASURES OF EYE-FIXATION PREDICTION

Several evaluation measures can be used for eye-fixation prediction, such as the normalized scanpath saliency (NSS), linear CC, information gain, area under the curve, earth mover's distance, similarity, Kullback–Leibler divergence (KLD), and receiver operating characteristic curve. From them, we selected four representative evaluation measures for evaluation: NSS, linear CC, area under the receiver operating characteristic curve (AUC), and KLD.

Let G and P represent the ground-truth and prediction maps, respectively. The NSS reflects the mean score assigned by the unit normalized saliency map, P_N, at human eye fixations. It was introduced by Peters and Itti [57] for saliency map evaluation and is defined as

$$NSS = \frac{1}{N}\sum_{i=1}^{N} P_N(i), \tag{9}$$

where N is the number of human eye fixations.

The linear CC measures the relationship between G and P, where scores of -1 or 1 indicate perfect linear relation between maps. The linear CC is given by

$$CC = \frac{\sigma(G,P)}{\sigma(G) * \sigma(P)}. \tag{10}$$

The AUC is widely used to evaluate maps estimated by saliency models. Two image locations are used to determine the AUC: a positive set of actual human fixations (i.e., fixation distribution) and negative set of points randomly sampled from the image (i.e., non-fixation distribution).

The KLD is based on information-theoretic measures to evaluate the information loss when a distribution is used to represent another approximate distribution. It corresponds to the probability interpretation of the eye-fixation prediction and the density map of the ground truth:

$$KLDiv = \sum_{i=1}^{N} G_i \log \frac{P_i}{G_i}. \tag{11}$$

D. ABLATION STUDY

The EGM provides edge information, and the FIM refines features and restores lost spatial information to improve the final saliency prediction map. To verify the effectiveness of these modules, we conducted an ablation study using the NCTU and NUS datasets. The results before and after removing the EGM and FIM are listed in Table 1 and shown in Fig.4. Furthermore, we also conducted an ablation study to show that the pyramid supervision training scheme effectively improves the proposed network performance. In conclusion, the proposed network (pyramid supervision) with EGM and FIM outperforms the variant without these modules.

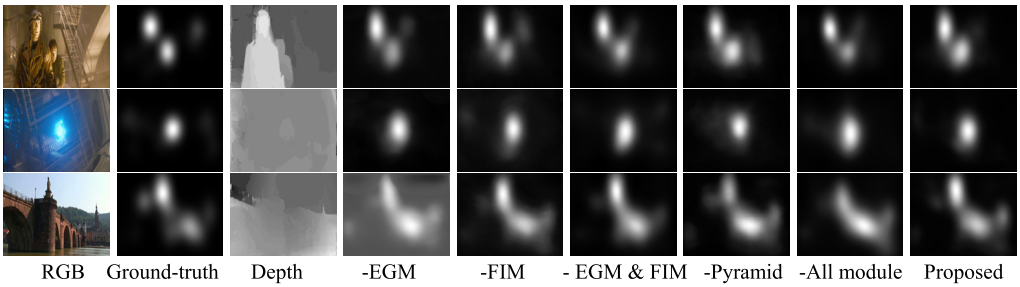

FIGURE 4. Examples of performance of proposed network with and without EGM and FIM.

TABLE 2. The effectiveness of asymmetric network.

Datasets	Criteria	Model A	Model B	Model C	Model D	Proposed
NCTU	CC	0.8040	0.8156	0.8189	0.8267	**0.8321**
	KLDiv	0.3120	0.3009	0.3027	0.3166	**0.2832**
	AUC	0.8728	0.8773	0.8753	0.8813	**0.8777**
	NSS	1.8688	1.9008	1.8994	1.9207	**1.9433**
NUS	CC	0.5356	0.5323	0.5326	0.5330	**0.5498**
	KLDiv	1.3012	1.3237	**1.1083**	1.3250	1.2251
	AUC	0.8518	0.8350	0.8521	0.8412	**0.8532**
	NSS	2.2017	2.2279	2.1150	2.2143	**2.2877**

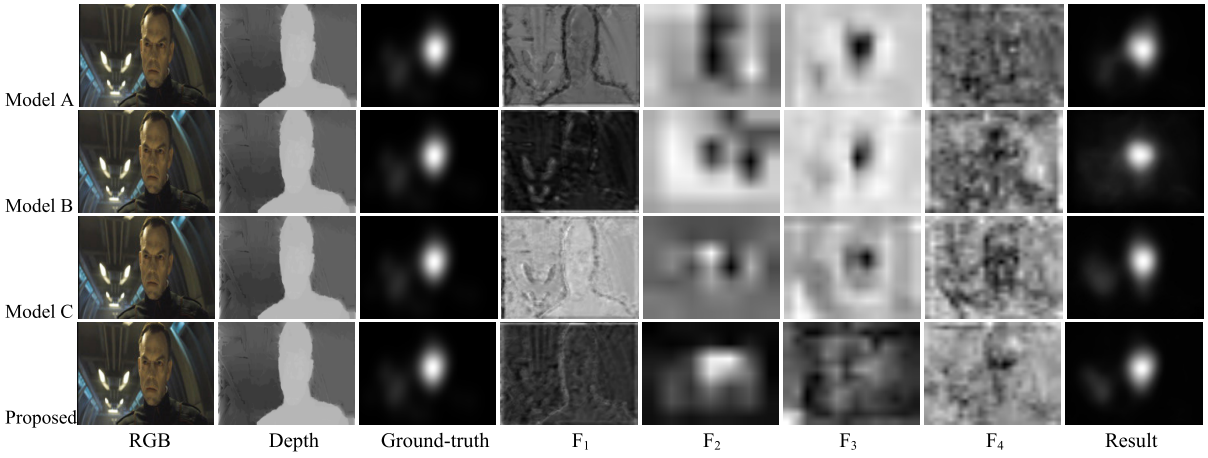

FIGURE 5. The visualization of the output of each module (Models A, B, and C, Proposed).

We observed that the current CNNs for the eye-fixation prediction are limited because CNNs used for the eye-fixation prediction are essentially a symmetric dual-stream input encoder structure. Although this same encoder structure improves the accuracy of the results, it also imposes a bottleneck on the eye-fixation prediction. To verify the effectiveness of asymmetric network in the proposed model, we performed another ablation experiment. For this analysis, three variations of the model were constructed. We devised three performance comparison models, namely, A, B, C and D. In Model A, we keep the pipeline unchanged while only replacing the ResNet34 with VGGNet16 (two streams of VGGNet, symmetric network). In Model B, we keep the pipeline unchanged while only replacing the VGGNet16 with ResNet34 (two streams of ResNet34, symmetric network). In Model C, we use ResNet34 as the encoding network for RGB images and VGGNet16 as the encoding network for depth maps. In Model D, we use VGGNet16 as the encoding network for RGB images and ResNet50 as the encoding network for depth maps. Table 2 summarizes the performances of models A, B, C, and our model. Figure 5 shows the features (visualization of the output of each module) generated by symmetric networks (models A and B) and the features captured by the asymmetric architectures (model C and Proposed model). The results of models A and B show that using the two asymmetric streams to extract features from RGB and depth information can improve the performance. The results of model C and the

TABLE 3. Performance of proposed and state-of-the-art methods on NCTU and NUS datasets.

Datasets	Criteria	GBVS	Fang	Qi	MLnet	DVA	DeepFix	iSEEL	EML-Net	Proposed
NCTU	CC	0.533	0.542	0.595	0.696	0.6834	**0.7974**	0.7578	0.7556	**0.8321**
	KLDiv	0.619	0.674	0.616	0.900	1.1045	1.3083	0.3985	**0.3886**	**0.2832**
	AUC	0.789	0.806	0.816	0.835	0.8023	0.8650	0.8315	**0.8818**	**0.8777**
	NSS	1.184	1.264	1.373	1.588	1.5546	1.8575	1.7187	**2.0666**	**1.9433**
NUS	CC	0.396	0.333	0.371	0.446	0.4549	0.4322	**0.5195**	**0.4857**	**0.5498**
	KLDiv	**1.374**	1.560	1.505	1.780	2.4349	1.8138	**1.2479**	2.2353	**1.2251**
	AUC	0.824	0.795	0.806	0.766	0.7236	0.7699	**0.8273**	**0.8149**	**0.8532**
	NSS	1.441	1.209	1.357	1.821	1.7962	1.6608	**2.1250**	**1.9419**	**2.2877**

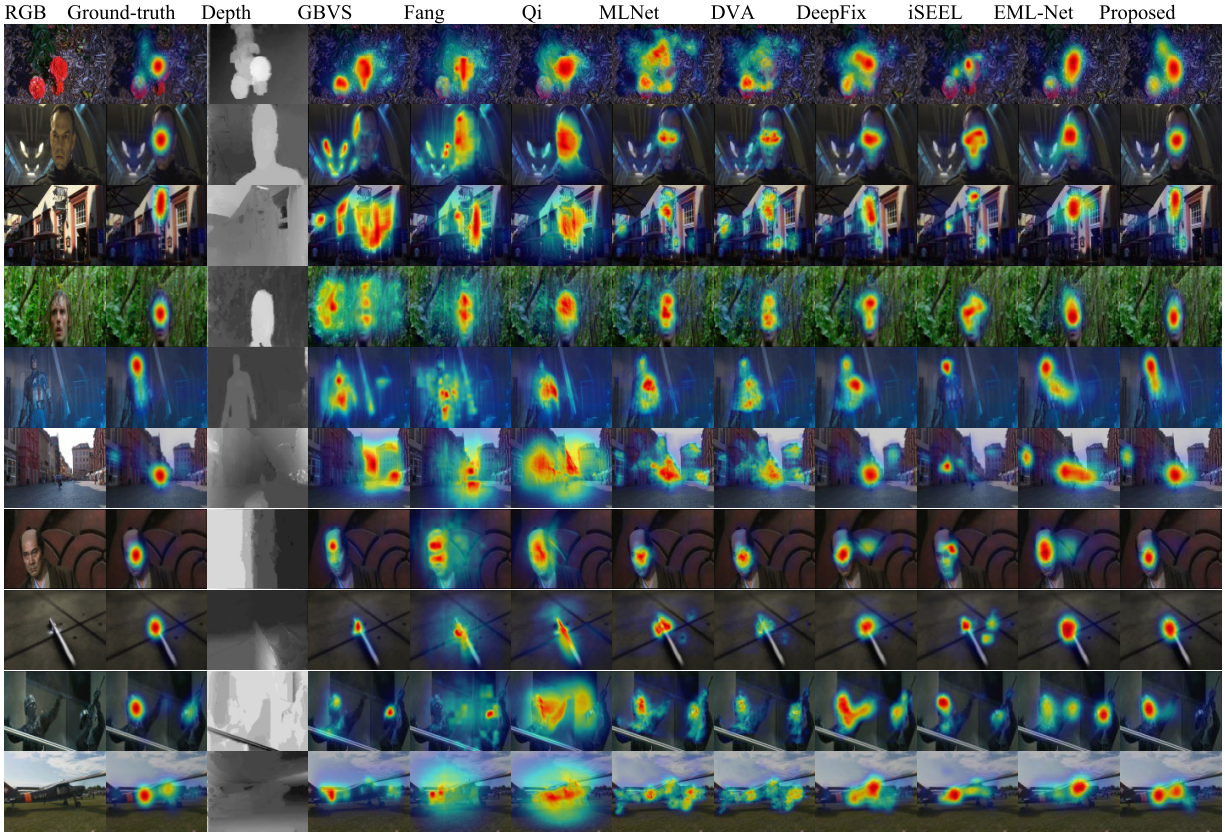

FIGURE 6. Examples of saliency maps obtained using the proposed network and other methods.

proposed method show that using VGGNet16 as the encoding network for RGB images and ResNet34 as the encoding network for depth maps can obtain better results.

E. PERFORMANCE EVALUATION

To demonstrate the performance of the proposed network to predict human eye fixation, its results were compared with those from eight state-of-the-art methods, including three traditional methods (GBVS [26], Qi [28], and Fang [58]), and five CNN-based methods (MLNet [34], EMLNet [35], DeepFix [36], DVA [59], and iSEEL [60]). The quantitative evaluation results are listed in Table 3. Our model does not obtain a big gain in performance especially on the AUC and NSS for the NTU dataset. This can be explained considering that the AUC is primarily based on true positives without significantly penalizing false positives. Although some methods may be effective on some evaluation measures, the proposed method is competitive with the best performing methods. In general, the proposed network achieves remarkable advantages over the comparison methods. To further illustrate the effectiveness of the proposed network, Figure 6 shows eye-fixation prediction maps

obtained from our network and the comparison CNN-based methods.

Figure 7 shows some failure cases generated by the proposed network in complex scenes. The EGM complements spatial features by acquiring edge information, and the FIM fuses features using the attention mechanism and cross-modal fusion. However, when edge information is not salient and the features are difficult to recognize, the prediction may substantially differ from the real eye fixation results. In future work, we will address these problems to further improve the proposed network.

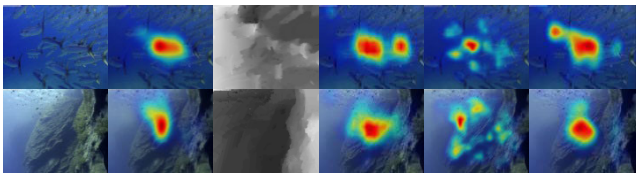

FIGURE 7. Examples of failure cases of proposed network and two state-of-the-art methods on images from NCTU dataset.

IV. CONCLUSION

We address two problems in human eye-fixation prediction using RGB-D data. One is the loss of spatial information during feature extraction, and the other is the proper use of the complementarity and correlation to fuse RGB and depth data. We propose an asymmetric network to solve these problems. The FIM refines input features and mines depth information from RGB and depth data after simple fusion. In addition, the EGM extracts the rich edge information in shallow network layers to complement the spatial information for improved decoding. As the RGB and depth modes contain different characteristics, we gradually perform a layered fusion during decoding to effectively preserve these characteristics and improve the eye-fixation predictions. Experimental results demonstrate that the proposed network contributes to the human eye-fixation prediction and can outperform existing state-of-the-art methods.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[2] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[3] W. Zhou, J. Yuan, J. Lei, and T. Luo, "TSNet: Three-stream self-attention network for RGB-D indoor semantic segmentation," *IEEE Intell. Syst.*, early access, Jun. 10, 2020, doi: 10.1109/MIS.2020.2999462.

[4] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[5] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "SG-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3855–3865, Sep. 2020, doi: 10.1109/TCYB.2020.2992433.

[6] S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[7] W. Zhou, J. Lei, Q. Jiang, L. Yu, and T. Luo, "Blind binocular visual quality predictor using deep fusion network," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 883–893, 2020.

[8] H. Hadizadeh and I. V. Bajic, "Saliency-aware video compression," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 19–33, Jan. 2014.

[9] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for HEVC-MSP," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 155–170, Jan. 2018.

[10] Y. Guo, F. Liu, J. Shi, Z.-H. Zhou, and M. Gleicher, "Image retargeting using mesh parametrization," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 856–867, Aug. 2009.

[11] J. Lei, M. Wu, C. Zhang, F. Wu, N. Ling, and C. Hou, "Depth-preserving stereo image retargeting based on pixel fusion," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1442–1453, Jul. 2017.

[12] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[13] T. Guo, J. Dong, H. Li, and Y. Gao, "Simple convolutional neural network on image calssification," in *IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, vol. 2017, pp. 721–724.

[14] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[15] Y. Gao, M. Shi, D. Tao, and C. Xu, "Database saliency for fast image retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 359–369, Mar. 2015.

[16] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image Retrieval–an empirical odyssey," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6439–6448.

[17] J. Wang, M. P. DaSilva, P. LeCallet, and V. Ricordel, "Computational model of stereoscopic 3D visual saliency," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2151–2165, Jun. 2013.

[18] H. Cheng, J. Zhang, Q. Wu, P. An, and Z. Liu, "Stereoscopic visual saliency prediction based on stereo contrast and stereo focus," *EURASIP J. Image and Vide Process*, vol. 61, pp. 1–13, Sep. 2017, doi: 10.1186/s13640- 017-0210-5.

[19] D. Liu and Z. Chen, "Disparity tuning guided stereoscopic saliency detection for eye fixation prediction," *J. Vis. Commun. Image*, vol. 57, pp. 218–227, Nov. 2018, doi: 10.1016/j.jvcir.2018.10.002.

[20] Y. Yang, B. Li, P. Li, and Q. Liu, "A two-stage clustering based 3D visual saliency model for dynamic scenarios," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 809–820, Apr. 2019.

[21] W. Zhou, J. Wu, J. Lei, J.-N. Hwang, and L. Yu, "Salient object detection in stereoscopic 3D images using a deep convolutional residual autoencoder," *IEEE Trans. Multimedia*, early access, Sep. 21, 2020, doi: 10.1109/TMM.2020.3025166.

[22] J. Wu, W. Zhou, T. Luo, L. Yu, and J. Lei, "Multiscale multilevel context and multimodal fusion for RGB-D salient object detection," *Signal Process.*, vol. 178, Jan. 2021, Art. no. 107766.

[23] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, "Going from RGB to RGBD saliency: A depth-guided transformation model," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3627–3639, Aug. 2020.

[24] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[25] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1448–1457.

[26] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 545–552.

[27] L. Zhang, Z. Gu, and H. Li, "SDSP: A novel saliency detection method by combining simple priors," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 171–175.

[28] F. Qi, D. Zhao, S. Liu, and X. Fan, "3D visual saliency detection model with generated disparity map," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 3087–3103, Jan. 2017.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional network for large-scale recognition," in *Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2015, pp. 1409–1556.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[32] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth maters:Influence of depth cues on visual saliency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2012, pp. 101–115.

[33] C.-Y. Ma and H.-M. Hang, "Learning-based saliency model with depth information," *J. Vis.*, vol. 15, no. 6, p. 19, May 2015.

[34] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3488–3493.

[35] S. Jia and N. D. Bruce, "EML-NET:An expandable multi-layer NETwork for saliency prediction," *Image Vis. Comput.* vol. 3, Feb. 2020, Art. no. 103887, doi: 10.1016/j.imavis. 2020. 103887.

[36] S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix:A fully convolutional neural network for predicting human eye fixations," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4456–44446, Sep. 2017.

[37] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2163–2176, Aug. 2020.

[38] Y. Lv, W. Zhou, J. Lei, and T. Luo, "Attention-based-based fusion network for human eye-fixation prediction in 3D images," *Opt. Express*, vol. 27, no. 23, pp. 34056–34066, 2019.

[39] H. Cheng, J. Zhang, Q. Wu, and P. An, "A computational model for stereoscopic visual saliency prediction," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 678–689, Mar. 2019.

[40] W. Zhou, Y. Lv, J. Lei, and L. Yu, "Global and local-contrast guides content-aware fusion for RGB-D saliency prediction," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Dec. 24, 2020, doi: 10.1109/TSMC.2019.2957386.

[41] A.-D. Nguyen, J. Kim, H. Oh, H. Kim, W. Lin, and S. Lee, "Deep visual saliency on stereoscopic images," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1939–1953, Apr. 2019.

[42] C. Xia, F. Qi, G. Shi, and C. Lin, "Stereoscopic saliency estimation with background priors based deep reconstruction," *Neurocomputing*, vol. 321, pp. 126–138, Dec. 2018.

[43] A. Banitalebi-Dehkordi, MT. Pourazad, and P. Nasiopoulos, "A learning-based visual saliency prediction model for stereo-scopic 3D video (LBVS-3D)," *Multimed. Tools Appl.*, vol. 76, no. 22, pp. 23859–23890, Nov. 2017, doi: 10.1007. /s11042-016-4155-y.

[44] Q. Zhang, X. Wang, J. Jiang, and L. Ma, "Deep learning features inspired saliency detection of 3D images," in *Proc. Pacificrim Conf. Multimedia*, 2016, pp. 580–589.

[45] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. C. H. Hoi, and H. Ling, "Learning unsupervised video object segmentation through visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3064–3074.

[46] M.-X. Jiang, C. Deng, J.-S. Shan, Y.-Y. Wang, Y.-J. Jia, and X. Sun, "Hierarchical multi-modal fusion FCN with attention model for RGB-D tracking," *Inf. Fusion*, vol. 50, pp. 1–8, Oct. 2019, doi: 10.1016/j.inffus.2018.09.014.

[47] B. Li, Q. Liu, X. Shi, and Y. Yang, "Graph-based saliency fusion with superpixel-level belief propagation for 3D fixation prediction," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2321–2325.

[48] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.

[49] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8554–8564.

[50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[51] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W.-S. Zheng, J. Li, and A. Wong, "Squeeze-and-attention networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13065–13074.

[52] Z. Zhang, Z. Lin, J. Xu, W. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for RGB-D salient object detection," 2020, *arXiv:2004.14582*. [Online]. Available: http://arxiv.org/abs/2004.14582

[53] Z. Zhang and Y. Pang, "CGNet:cross-guidance network for semantic segmentation," *Sci. China Inf. Sci.*, vol. 63, no. 2, Feb. 2020, Art. no. 120104, doi: 10.1007/s11432-019-2718-7.

[54] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3060.

[55] R. Huang, Y. Xing, and Z. Wang, "RGB-D salient object detection by a CNN with multiple layers fusion," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 552–556, Apr. 2019.

[56] H. Chen, Y. Li, and D. Su, "Mutil-modal fusion network with mutil-scale mutil-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 15, pp. 367–385, Dec. 2019.

[57] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[58] Y. Fang, J. Lei, J. Li, L. Xu, W. Lin, and P. L. Callet, "Learning visual saliency from human fixations for stereoscopic images," *Neurocomputing*, vol. 266, pp. 284–292, Nov. 2017.

[59] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[60] H. R. Tavakoli, A. Borji, J. Laaksonen, and E. Rahtu, "Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features," *Neurocomputing*, vol. 244, pp. 10–18, Jun. 2017.

WENYU LIU received the B.S. degree from the School of Information Science and Technology, Nantong University, Nantong, China, in 2018. She is currently pursuing the M.S. degree with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Zhejiang, China. Her research interests include multimedia signal processing and communication.

WUJIE ZHOU (Member, IEEE) is currently an Associate Professor with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Zhejiang, China. He is also a Postdoctoral Fellow with the Institute of Information and Communication Engineering, Zhejiang University, Zhejiang. His research interests include multimedia signal processing and communication. He is the Reviewer of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON BROADCASTING, the IEEE SIGNAL PROCESSING LETTERS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: EXPRESS BRIEFS, *Information Sciences*, *Neurocomputing*, and *SPIC*.

TING LUO received the B.S. degree in computer science from Ningbo University, Ningbo, China, in 2003, the M.S. degree in business information technology from Middlesex University, London, U.K., in 2004, and the Ph.D. degree from Ningbo University, in 2016. He is currently a Professor with the College of Science and Technology, Ningbo University, Zhejiang, China. His research interests include multimedia security, image processing, data hiding, and pattern recognition.