IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 9, 2020, accepted January 20, 2020, date of publication February 4, 2020, date of current version February 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2971509

Multi-Modal Weights Sharing and Hierarchical Feature Fusion for RGBD Salient Object Detection

FEN XIAO[1], BIN LI[1], YIMU PENG[1], CHUNHONG CAO[1], KAI HU[1], AND XIEPING GAO[1,2], (Member, IEEE)

[1]Key Laboratory of Intelligent Computing and Information Processing, Ministry of Education, Xiangtan University, Xiangtan 411105, China
[2]College of Software and Communication Engineering, Xiangnan University, Chenzhou 423043, China

Corresponding authors: Xieping Gao (xpgao@xtu.edu.cn)

This work was supported in part by the China Education and Research Network (CERNET) Innovation Project under Grant NGII20180309, the National Natural Science Foundation of China under Grant 61972333 and Grant 61802328, the Natural Science Foundation of Hunan province in China under Grant 2018JJ2405, and the Scientific Research Fund of Hunan Provincial Education Department under Grant 18K034.

ABSTRACT Salient object detection (SOD) aims to identify and locate the most attractive regions in an image, which has been widely used in various vision tasks. Recent years, with the development of RGBD sensor technology, depth information of scenes becomes available for image understanding. In this paper, we systematically investigate and evaluate on how to integrate depth cues in a pre-trained deep network and learn informative features for SOD. First, we propose a CNN-based cross-modal transfer learning, which learn knowledge from sufficient labeled RGB salient object datasets and guide the depth domain feature extraction. Then we design a feature fusion module to fuse the complementary features in a hierarchical manner. At last, the final saliency map is obtained by integrating multi-scale information step by step. Extensive experiments on five popular RGBD benchmark datasets demonstrate that our proposed approach achieves significant improvements and outperforms the state-of-the-art methods.

INDEX TERMS RGBD, salient object detection, complementary feature extraction, hierarchical fusion.

I. INTRODUCTION

SOD aims at capturing the most visually distinctive objects and informative regions, which attract human more attention in a scene. Many computer vision applications may benefit from understanding where humans interest in an image, such as object detection [1], [2], content-aware image editing [3], image compression [4], [5], image retrieval [6] and image synthesis [7].

Saliency detection usually involved into top-down and bottom-up algorithms. Top-down approaches are task-driven which normally is a combination of insights from neuro-sciences, biology and computer vision. While bottom-up approaches are stimuli-driven, which by aggregating low-level image features, such as color, edge and shape, to detect salient regions in a scene. Based on above two frameworks, various RGB-domain SOD methods have been proposed and achieved remarkable performance. However, they may fail to detect salient object in some complex scenes, such as the similar color distribution of objects and backgrounds or varying illuminations. Recently, low-cost RGBD sensors (Microsoft Kinect [8] or Intel RealSense) have become very popular to capture both color images and depth maps for a scene at the same time. Depth cues contain many correctly object structures, which have shown its efficiency for salient object segmentation [9]–[11]. However, the issue of how to effectively utilize the depth and RGB information remains to be solved.

There are many methods focus on designing handcrafted features from RGBD data with domain-specific knowledge [12]–[14]. Normally, RGB and depth features are integrated in a heuristic way, which fail to take full advantage of complementary information to describe complex correlations. Furthermore, the handcrafting process is highly dependent on prior knowledge of existing datasets, and resulting features may be intuitive but do not work well enough for other case. Recently, convolutional neural network (CNN) based image feature extraction has attracted many attentions

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

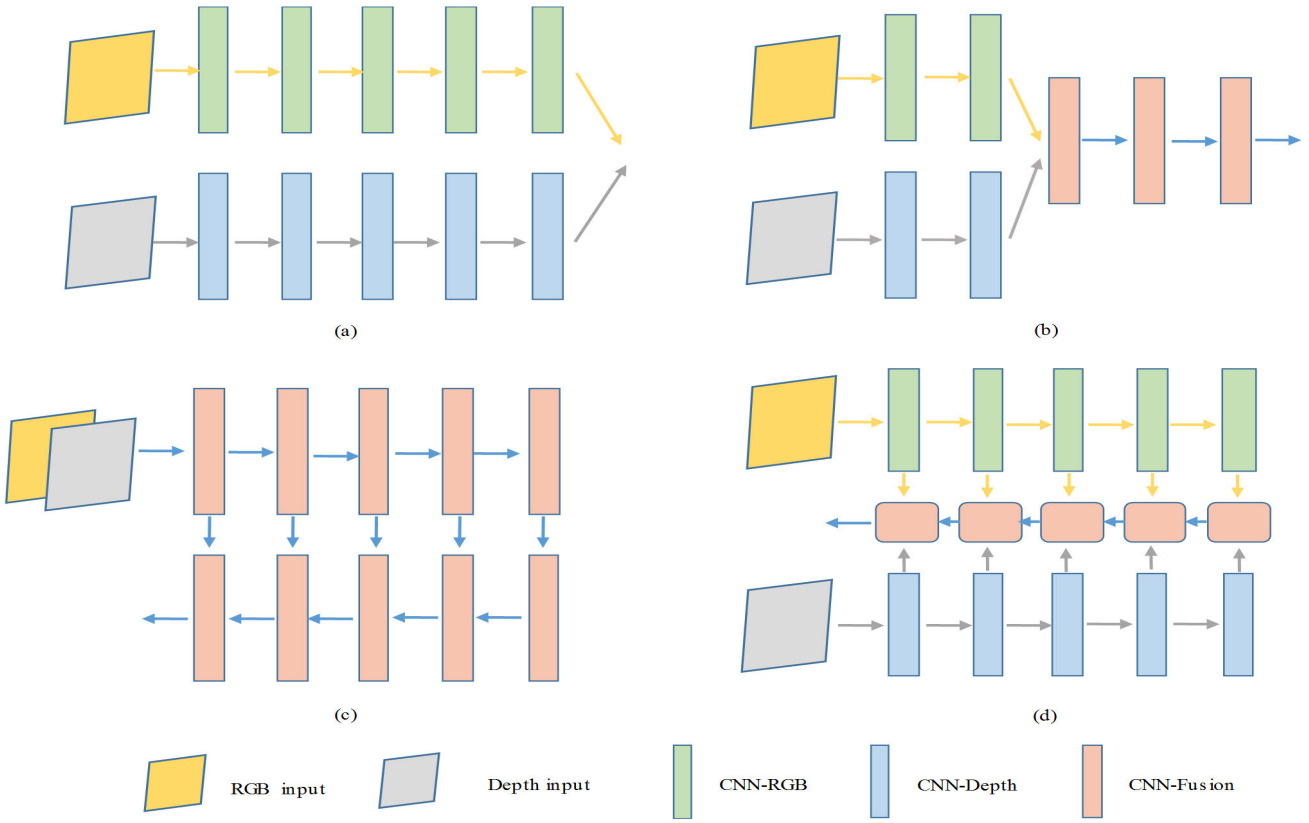


FIGURE 1. Different network structures for deep feature fusion. (a) Late fusion. (b) Early fusion. (c) Ad-multi-fusion. (d) Hierarchical fusion.

of researchers. An obvious advantage of this method is that can establish an effective hierarchical representation of the input data, which integrates RGB features in a convolutional way. Also, many deep learning models have been adopted for dealing RGBD data [15]–[20] and achieved pretty well performance.

Many works show that the performance of CNN-based SOD is closely related on the number of layers and the volume of labeled training datasets. Recently, owing to the emergence of extensive annotate datasets, such as MSRA10K, DUTS and ECSSD, many efficient models have been proposed, which can give satisfactory results for RGB salient object detection. While for RGBD image processing, it is impractical to build RGBD benchmarks to match the scale and variety of existing RGB datasets. Some methods [21], [22], utilized the knowledge learned from large-scale RGB datasets to guide the RGBD training stage, have achieved a great success. These works generally follow encoder-decoder architecture that focus on the cross-modal feature extraction and complementary feature integration. As shown in Fig.1, we illustrate four fusion models in which fuse RGB and depth information on different stage. In Fig.1(a), RGB and depth data are processed independently during the encoding-decoding stage, then simply fuse the outputs of each stream to get final saliency map. The early fusion structure is illustrated in Fig.1(b), which encodes RGB and depth channel separately and then integrates the feature maps of both channels before decoding. Finally, the fused feature vector can be decoded to get the saliency map. Fig.1(c) shows the Ad-multi-fusion module, proposed by [22], in which RGBD four channel feeds in a network and encodes as a concatenated multi-scale feature vector, the de-convolutional layers then decode spatial features using the feature maps of the previous layer to get more details of salient regions. The above methods combine RGB and depth modalities on a single scale, which may not be sufficient to obtain enough useful information from the complementary modality. Fig.1(d) gives a hierarchical fusion strategy, during the encoding stage, the RGB and depth information are extracted in a hierarchically manner. While in the decoding stage, both RGB and depth channel feature maps in the same scale are fused and fed into the decoder, and then progressively decodes layer by layer to get the saliency map. In this manner, we can locate more salient objects with large receptive fields and maintain more context information from the upper layer to the lower layer.

In this paper, we propose a new pre-trained guide hierarchical fusion network (PGHF) for RGBD salient object detection. The PGHF consist of two VGG16 based feature extraction branches and a hierarchical fusion CNN branch (HF-CNN). Firstly, both RGB and depth information are jointly extracted by two VGG 16 feature extraction branches.

Two networks share the same weights, which pre-trained on a large-scale RGB dataset as the initial parameter of the feature extraction layer. Then, the HF-CNN is used to fuse the multi-modal features in a hierarchical manner and produce multi-scale coarse saliency maps progressively. At last, based on multi-scale coarse saliency maps, we achieve final estimation of salient objects.

The contributions of our work are summarized as follows:

1) We built a novel network, i.e., PGHF, which is specifically designed to learn complementary RGBD features and can detect salient regions more accurately under the guidance of pre-trained RGB model.

2) We propose a hierarchical fusion architecture, which can leverage semantic relations and multi-scale information between different layers. Algorithm outputs more detailed object boundaries and keeps different scale salient objects to be spatially consistent.

3) Extensive experimental results on five popular RGB-D saliency benchmarks demonstrate that the proposed approach outperforms the state-of-the-art methods in terms of both regional and boundary evaluation measures.

II. RELATED WORK

A. RGB SALIENCY DETECTION

Over the past decades, many saliency object detection approaches for RGB image have been developed. Early methods detect salient objects based on some pre-designed features. Achantay et al. [23] presented a frequency-tuned approach utilize low level manual features of color and luminance to computing saliency in images. Cuong and Chandler [24] employed lightness distance, color distance, contrast, sharpness and edge strength, which estimate the validity of each feature map based on local measure of cluster density with image-adaptive technique. Cheng *et al.* [25] proposed saliency extraction algorithm based on regional contrast, which simultaneously evaluates global contrast differences and spatial coherence. Niu *et al.* [26] developed superpixel and background connectivity prior to achieve salient object segmentation. These methods have achieved massive development on saliency detection. However, they are limited to hand-crafted saliency cues, which make them the low representation capability of high-level relations and the high generalization errors. Inspired by the successful application of deep convolutional networks in various computer vision tasks for high-quality representations, these network structures are also adopted to RGB saliency detection with more informative representations. Li and Yu [27] designed deep CNNs with multiple fully connected layers to extract multi-scale features further boost the performance for salient object detection. Yi *et al.* [28] presented a top-down structure with horizontal connections for constructing high-level semantic feature maps of various scales. Liu and Han [29] introduced a hierarchical architecture which supervise the multi-scale coarse saliency maps with different scales ground truth maps to refine the details of saliency map hierarchically and progressively. Zhuge *et al.* [30] provided stage-wise refinement frameworks to gradually enhance the boundary information. These methods make remarkable success. However, it is still difficult to distinguish salient object from complex image scenes.

B. RGBD SALIENCY DETECTION

Recently RGBD sensors not only provide RGB information but also record the corresponding depth information. Depth has been shown to be one of the practical cues for predicting saliency. Ciptadi *et al.* [31] extracted layout and shape features from the depth cues to improve the ability of salient object detection. Fan *et al.* [32] combined depth weighted color contrast and spatial compactness to generate saliency object maps. Ren *et al.* [33] exploited depth global priors to estimated saliency maps. Ju *et al.* [12] proposed a depth-aware method for saliency detection using an anisotropic center-surround difference measure. Feng *et al.* [13] defined depth background enclosure to detect salient objects in RGB-D images. Zhu *et al.* [34] utilized center-dark channel prior of depth maps to innovate the salient object detection. Qu et al. [35] proposed a CNN to automatically learn the low-level saliency depth cues and combine the Laplacian propagation framework with the convolutional neural network to produce spatially consistent saliency map. Han *et al.* [36] pointed out the difference between RGB and depth data, and designed a cross-modal transfer learning CNN architecture to better extract the modal-specific features depth cues. Chen and Li [37] presented a hierarchical fusion module for depth fusion, which using cross-modal residual function and jointly supervision to learn the complementary features from the paired modality dataset. In [38], Zhao *et al.* introduced a contrast-enhanced net for depth images and built a fluid pyramid integration module to make better use of multi-scale cross-modal features. However, due to the limited labeled training datasets, it may be difficult to learn effective feature representations from RGBD date via the CNNs. In order to solve these problems, Zhu *et al.* [21] designed a prior-model guided depth-enhanced network (PDNet) for salient object detection. They utilized a subnet to extract depth cues and then incorporates depth-based features into the primary RGB network. Finally, they employed a large RGB dataset to pre-train the master network and achieved significant accuracy improvement. Huang *et al.* [22] designed a CNN with multiple layer fusion (CMLF), which encodes RGBD four channel as a concatenated multi-scale feature vector, and then decode spatial features using the feature maps of the previous layer to get more details of salient regions. For dealing with the problem of insufficient datasets, they use zero maps to replace the depth map during the pretraining process for facilitating the prediction of RGBD.

III. THE PROPOSED METHOD

Our proposed model PGHF follows the encoder-decoder architecture as shown in Fig. 2. The encoder part consists of two complementary feature extraction branches $\mathrm{VGG}_{\mathrm{RGB}}$ and $\mathrm{VGG}_{\mathrm{Depth}}$ for cross-modal feature extraction. As for

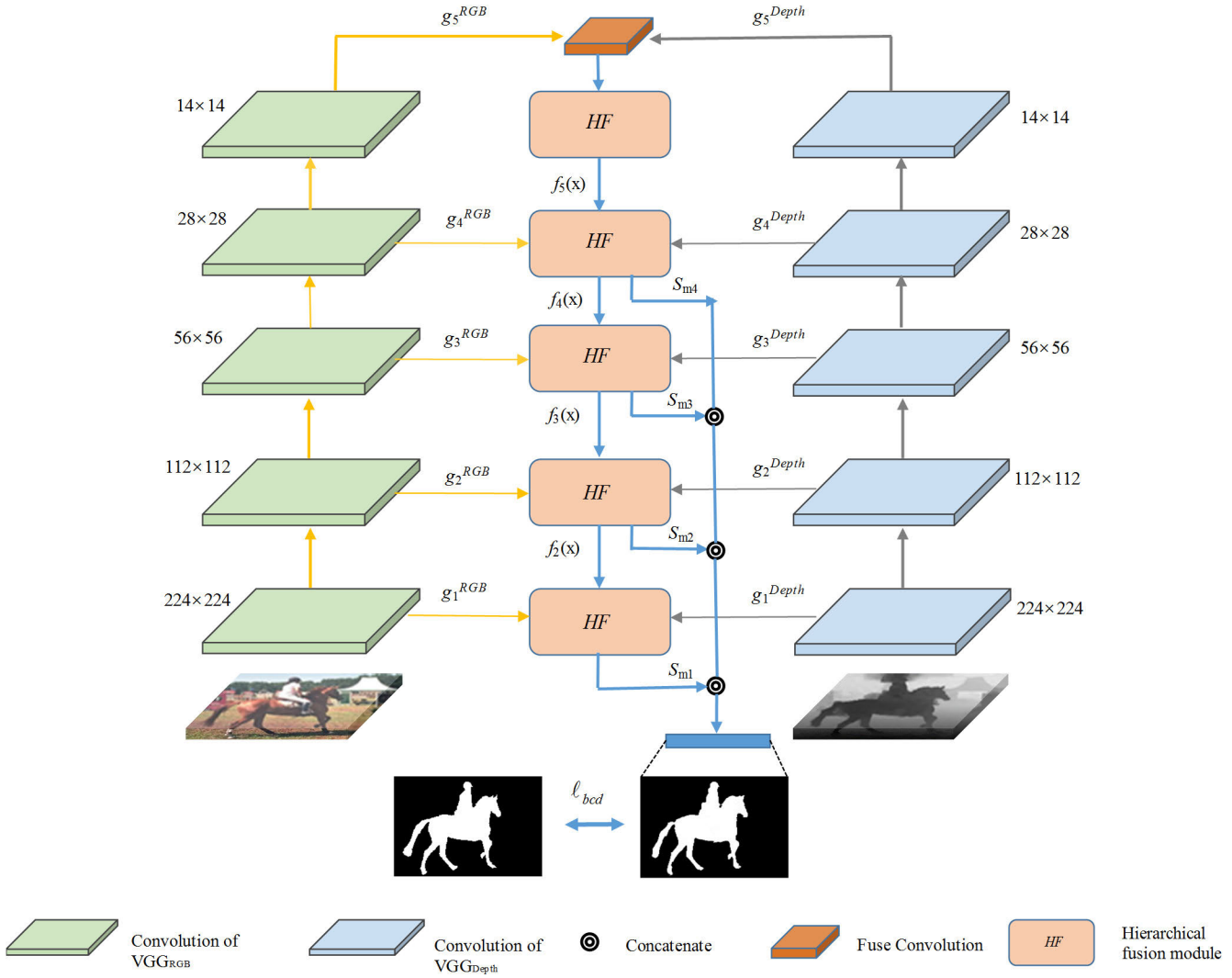


FIGURE 2. The architecture of our proposed PGHF model consists of three deep networks. Both left and right sides of our model are feature extractors named as VGG_{RGB} and VGG_{Depth} respectively. The middle column is a fusing branch, which hierarchically integrates the complementary features and outputs multi-scale saliency maps progressively.

the decoder, we employ a hierarchical fusing CNN branch (HF-CNN) to fuse multi-scale features and finalize the salient map estimating. In the following subsections, we will describe the details of the complete architecture and possible advantage of extracting useful information for saliency prediction.

A. COMPLEMENTARY FEATURE EXTRACTION

As shown in Fig. 2, VGG_{RGB} and VGG_{Depth} are based on the VGG-16 model which has 13 convolutional layers (kernel size $= 3 \times 3$, stride $= 1$) and 4 maxpooling layers (pooling size $= 2 \times 2$, stride $= 2$). Given input RGB image X_{RGB} and its corresponding depth image X_{Depth}, we adopt the convolutional layers of VGG_{RGB} and VGG_{Depth} to extract five different scales complementary feature maps as the representations of the input image pair. For notation simplicity, we refer to the VGG as a function $g(X, \theta)$, which take X as input and θ as parameters. The complementary feature extraction process is denoted as

$$[g_l^{RGB}, g_l^{Depth}] = [g_l(X_{RGB}; \theta_{ws}), g_l(X_{Depth}; \theta_{ws})], \quad l = 1, 2, 3, 4, 5 \tag{1}$$

where g_l^{RGB} and g_l^{Depth} are the l-layer feature representation of images X_{RGB} and X_{Depth}, respectively. $[\cdot, \cdot]$ is the concatenation operator. θ_{ws} is the parameter of the convolutional layers in VGG_{RGB}. In our framework, the weights of both VGG_{Depth} and VGG_{RGB} are set same as pre-trained network and then be trained simultaneously.

B. HIERARCHICAL FUSION MODULE

We use the HF-CNN to integrate the multi-level complementary features of image pairs in a hierarchical manner. The hierarchical fusing function is defined by

$$f_l(X) = \begin{cases} h([g_l^{RGB}, g_l^{Depth}]; \theta_f), & l = 5 \\ h([g_l^{RGB} + g_l^{Depth}, f_{l+1}(X)]; \theta_f), & l < 5 \end{cases} \tag{2}$$

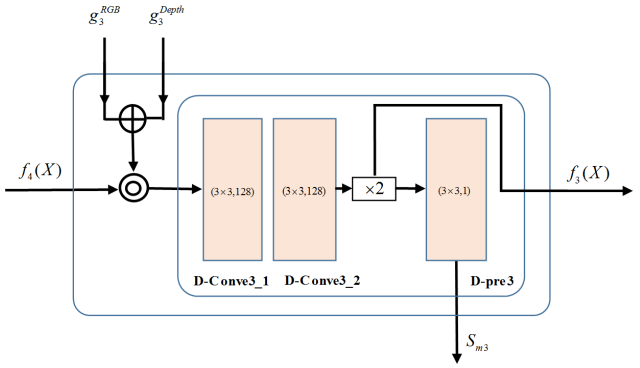


FIGURE 3. An illustration of hierarchical feature fusion in the third layer of HF-CNN.

where h denotes the feature fusion operator, θ_f is the parameter of the fusion branch. From the Eq. (2), we can see that the integration scheme is vary in different layers. In the last layer, we use only g_5^{RGB} and g_5^{Depth} as the input of the fifth layer fusion module. While for other layers, both the outputs of the previous fusion layer $f_{l+1}(X)$ and the sum of feature vectors, extracted by VGG_{RGB} and $\text{VGG}_{\text{Depth}}$, are taken into consideration.

For each decoder layer, we also output the coarse salient maps for final fusion. We upsample the fusion feature map $f_l(X)$ using a Bilinear interpolation and then apply a convolution network with a $3\times3\times1$ kernel to get the corresponding coarse salient map $S_{m,l}$:

$$S_{m,l} = Conv(U_0(f_l(X); o_l)), \quad l = 4, 3, 2, 1 \tag{3}$$

where U_0 denotes upsample operator, 2^l is the kernel with stride of 2^l. As an example, we give the fusion details of the third layer of hierarchical fusion module, shown in Fig.3.

Then we stack four scale coarse saliency maps together and convert the concatenated map into the final salient map with the following activation function:

$$S = \sigma([S_1, S_2, S_3, S_4]) \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function.

C. TRAIN

Given a set of samples $C = \{X_n, G_n\}_{n=1}^N$ with N training pairs, where $X_n = \{X_{RGB}, X_{Depth}\}$ are the input image pairs and G_n is the corresponding binary ground-truth image, $g_n^i = 1$ presents pixel i of image n is a saliency pixel, $g_n^i = 0$ denotes the background pixel. We use G_+ and G_- to denote pixel sets of salient region and background in the ground truth image, respectively.

For training, all the parameters are learnable via minimizing a loss function, which is computed as the errors between the final RGBD saliency maps and ground-truth images. To obtain high quality regional segmentation and clear boundaries, a hybrid loss is defined as:

$$\ell_{bcd} = \arg\min \mu\ell_{bc} + \lambda\ell_{dc} \tag{5}$$

where μ and λ are used to balance two loss functions and make two tasks achieve best performance. ℓ_{bc} and ℓ_{dc} denotes Binary Cross-Entropy (BCE) loss [40] and Dice-score loss [41], the specific descriptions are in follows.

BCE loss measures both structural and global difference between the predicted saliency map and the ground truth. The loss is defined as:

$$\ell_{bc} = -\sum_{i\in G_+} \log P(g_i=1|X;\theta) - \sum_{i\in G_-} \log P(g_i=0|X;\theta) \tag{6}$$

where $\theta = [\theta_{ws}, \theta_f]$ is the parameter in the entire network. $P(g_i = 1|X;\theta) \in [0, 1]$ indicates the predicted probability that pixel i to be a salient:

$$P(g_i = 1|X;\theta) = \frac{e^{s_i}}{e^{1-s_i} + e^{s_i}} \tag{7}$$

where s_i is the output of network.

Dice-score loss was first proposed for supervision medical image segmentation. It can alleviate the problem of class imbalance and local spatial inconsistency, which is defined as:

$$\ell_{cd} = \frac{2\sum_{i\in G_+} P(g_i = 1|X;\theta)}{|G_+| + \sum_{i\in G} P(g_i = 1|X;\theta)} \tag{8}$$

where $|G_+|$ denote the salient pixel numbers.

In the training phase, we pre-trained VGG16 network with 20553 labeled RGB images come from MSRA10K and DUTS-TR. Within the pre-training stage, only R, G, B three channels data are available and we train the VGG_{RGB} by minimizing the loss function and obtained the network parameters w. For the PGHF training, w is used as the initial parameter of both VGG_{RGB} and $\text{VGG}_{\text{Depth}}$, then we can fine tune the network parameters w with the supervision of RGBD salient object dataset. In this way, the representations learned from RGB data is transferred to the RGBD field.

IV. EXPERIMENTS

A. DATASETS

Five popular RGBD datasets NJUD2K [12], NLPR [42], SSB1000 [11], LFSD [43] and RGBD135 [44] are used for our experimental comparison. NJUD2K consists of 2003 stereo-scopic images, which are collected from the Internet, 3D movies, and photographs taken with a stereo-scopic camera. NLPR contains 1000 image pairs, which were captured by Kinect from indoor and outdoor scenes. SSB1000 comprises of 1000 pairs of binocular images from the available Web links. LFSD includes 100 images with depth information, which were taken via the Lytro light field camera [45], and their corresponding hand-labeled ground truth images. RGBD135 is composed of 135 indoor images, which collected by Microsoft Kinect of seven indoor scenes.

Following [38], the training set contains 2050 samples, which are consists of 1400 images from the NJU2K and 650 images from the NLPR. The validation set includes 150 samples, which combined with 100 images from NJU2K

TABLE 1. The results for component analysis on NJUD2K.

Inputs	Model	S-measure ↑	meanF ↑	maxF ↑	MAE ↓
RGB	VGG	0.809	0.796	0.808	0.086
Depth	VGG	0.735	0.695	0.709	0.132
RGBD	VGG+ Late-fusion	0.870	0.860	0.876	0.063
	VGG+ Early-fusion (PDNet[21])	0.874	0.814	0.872	0.076
	VGG+Ad-muti-fusion (CMLF[22])	--	0.867	--	0.056
	VGG+HF	0.874	0.870	0.880	0.052
	VGG+*ws*+HF (PGHF)	**0.882**	**0.876**	**0.884**	**0.050**

and 50 images from NLPR. And validate all comparison models on the rest images. All of the samples are randomly selected.

B. IMPLEMETATION DETAILS

We implement our method using Keras 2.2.4 on a PC with E5-2630 CPU and Nvidia Tesla GPU (with 16G memory). The initial learning rate is set to 0.0002, weight decay, momentum and batch size are set as 0.0005, 0.9 and 4. Training process converges after 25 epochs. The inference time for each 224 × 224 image pair is merely 0.042s.

C. EVALUATION METRICS

Four widely used metrics, including structural measure [46] (S-measure), mean F-measure [47] (meanF), maximum F-measure (maxF) and mean absolute error [48] (MAE), are used to evaluate the proposed method.

S-measure focuses on evaluating the similarity structure between the predicted saliency map and binary ground-truth, formulated as:

$$S = \alpha * S_o + (1 - \alpha) * S_r \tag{9}$$

where $\alpha \in [0, 1]$ is the balance parameter and be empirically set to 0.5. S_o and S_r is the object-aware and region-aware structural similarity between generated saliency map and the corresponding ground truth map.

F-measure is a harmonic mean of average precision and average recall, defined as:

$$F_\beta = \frac{(1 + \beta^2)Precision * Recall}{\beta^2 * Precision + Recall} \tag{10}$$

As suggested by [47], β^2 is set to be 0.3 for emphasizing the importance of precision. Following [49], we provide the max F-measure and mean F-measure by using different thresholds ranging from 0 to 255.

MAE represents the average pixel-wise absolute difference between an estimated saliency map and its corresponding ground truth. computed as:

$$MAE = \frac{1}{T} |s_i - g_i| \tag{11}$$

where T denotes the total number of pixels.

D. ABLATION STUDIES

In this section, we evaluate the performance of some parts in the proposed model. Table 1 gives salient object detection results of the architecture with different input channel, with/no weight sharing and four fusion schemes. Due to the limitation of space, we only show the results on NJUD2K dataset. As show in table 1, we use three different input data, which are RGB, Depth, and RGB + Depth. All of models given in this table are pre-trained with RGB SOD datasets. As for dealing with RGBD input, parameters of VGG-depth can be set with two ways: one is random, another is ws. ws means weight sharing, i.e. the network parameters of both VGG for RGB and depth feature extraction are same. The four fusion schemes Late-fusion, Early-fusion, Ad-muti-fusion and HF are illustrated in Fig.1.

TABLE 2. The results of PGHF model training with different losses on NJUD2K.

Model	Smeasure ↑	meanF ↑	maxF ↑	MAE ↓
PGHF- ℓ_{bc}	0.878	0.853	0.874	0.066
PGHF- ℓ_{cd}	0.872	0.868	0.876	0.053
PGHF- ℓ_{bcd}	0.882	0.876	0.884	0.050

1) EFFECT OF COMPLEMENTARY INPUTS

To investigate the impact of different inputs, we set up three date types: RGB, Depth and RGB + Depth. For only RGB(Depth) input, our model do not fuse different channel input information, which became a simple multi-level feature model. As the results shown in Table 1, we observe that only with the RGB(Depth), our model achieves 0.809(0.735), 0.796(0.695), 0.808(0.709) and 0.086 (0.132) in terms of the S-measure, meanF, maxF and S-measure metrics. Compared rows 1st and 3rd, we find that the additional depth information can upgrade the performance on all metrics, such as about 5.1% of S-measure, around 3.1% of meanF, near 4.6% of maxF, as well as around 10.5% decrease in MAE. It indicates the complementary effects of RGB and Depth are very expressive. Depth information enhances the spatial similarity of saliency maps, improves the accuracy of saliency object detection.

2) EFFECT OF SHARE WEIGHTS

As described in section III, we learn RGB salient object representations from massive labeled RGB images, and then use the same convolution layer weights to extract depth feature. As shown in Table 1, comparing the 6^{th} and 7^{th} rows, we could see that the performance

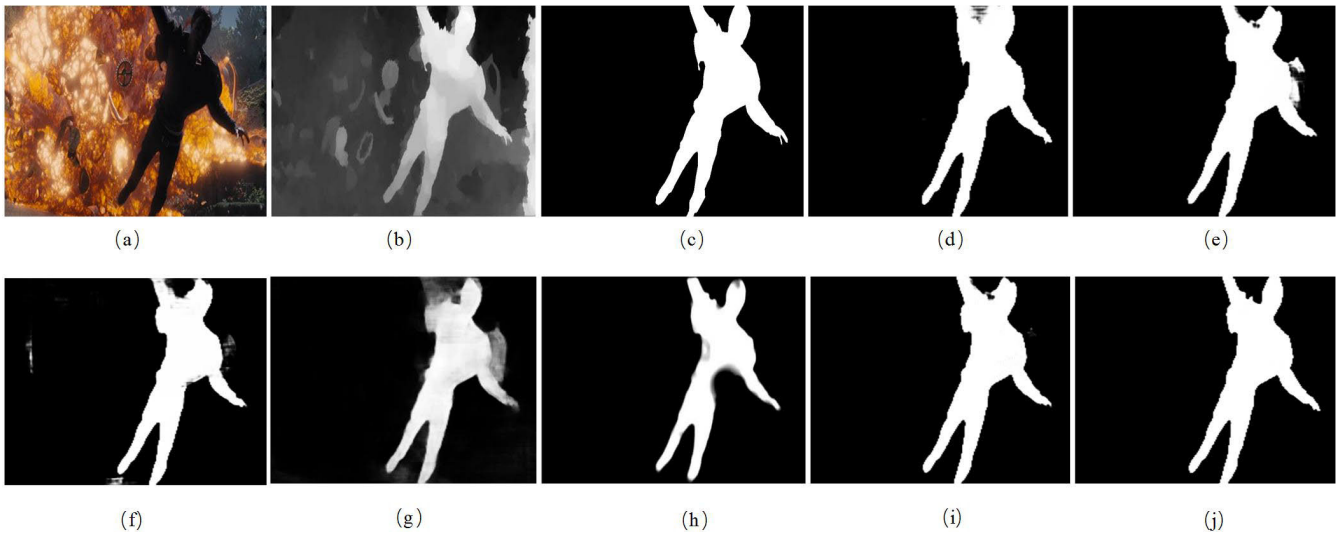


FIGURE 4. Visualization of typical saliency maps with different component. (a) RGB. (b) Depth. (c) Ground truth. (d) VGG_{RGB}. (e) VGG_{Depth}. (f) VGG + Late-fusion. (g) VGG + Early -fusion. (h) VGG + Ad-muti-fusion. (i) VGG + HF. (j) VGG +*ws*+ HF (PGHF).

TABLE 3. Quantitative comparison with 6 methods on five popular RGB-D saliency benchmarks.

Dataset	Metirc	CDCP *ICCVW* 2017	DF *TIP* 2017	CTMF *Trans 2017*	PDNet *ICME 2019*	PCF *CVPR* 2018	CPFP *CVPR* 2019	PGHF *OURS*
SSB1000	S-measure ↑	0.713	0.757	0.848	0.874	0.875	0.879	**0.883**
	meanF ↑	0.643	0.616	0.758	0.814	0.818	0.842	**0.860**
	maxF ↑	0.668	0.756	0.831	0.872	0.860	0.873	**0.880**
	MAE ↓	0.149	0.141	0.086	0.076	0.064	0.051	**0.047**
NJU2000	S-measure ↑	0.669	0.763	0.849	0.874	0.877	0.878	**0.882**
	meanF ↑	0.594	0.663	0.779	0.814	0.840	0.850	**0.876**
	maxF ↑	0.621	0.815	0.845	0.872	0.872	0.877	**0.884**
	MAE ↓	0.180	0.136	0.085	0.076	0.059	0.053	**0.050**
LFSD	S-measure ↑	0.717	0.791	0.796	0.832	0.794	0.828	**0.835**
	meanF ↑	0.680	0.679	0.756	0.738	0.761	0.811	**0.835**
	maxF ↑	0.703	0.817	0.791	0.826	0.779	0.826	**0.840**
	MAE ↓	0.167	0.138	0.119	0.126	0.112	0.088	**0.083**
RGBD135	S-measure ↑	0.709	0.752	0.863	0.884	0.842	0.872	**0.923**
	meanF ↑	0.585	0.604	0.756	0.776	0.765	0.815	**0.907**
	maxF ↑	0.631	0.766	0.844	0.865	0.804	0.838	**0.919**
	MAE ↓	0.115	0.093	0.055	0.056	0.049	0.037	**0.021**
NLPR	S-measure ↑	0.727	0.802	0.860	0.878	0.874	0.888	**0.909**
	meanF ↑	0.621	0.684	0.753	0.771	0.809	0.840	**0.883**
	maxF ↑	0.655	0.792	0.834	0.869	0.847	0.869	**0.897**
	MAE ↓	0.117	0.078	0.063	0.057	0.052	0.036	**0.027**

of the model has been greatly improved with weights sharing.

3) EFFECT OF HIERARCHICAL FUSION MODULE

We report the quantitative comparison results of different fusion modules in 3rd-6th rows of the Table 1. Compare with other fusion strategies, the hierarchical fusion module achieves optimal performance on F-measure and MAE. The main reason is that the proposed fusion module can maintain more context information from the upper layer to the lower layer. Visual examples for different fusion strategies are display in Fig. 4 (f)-(i). We can observe that hierarchical

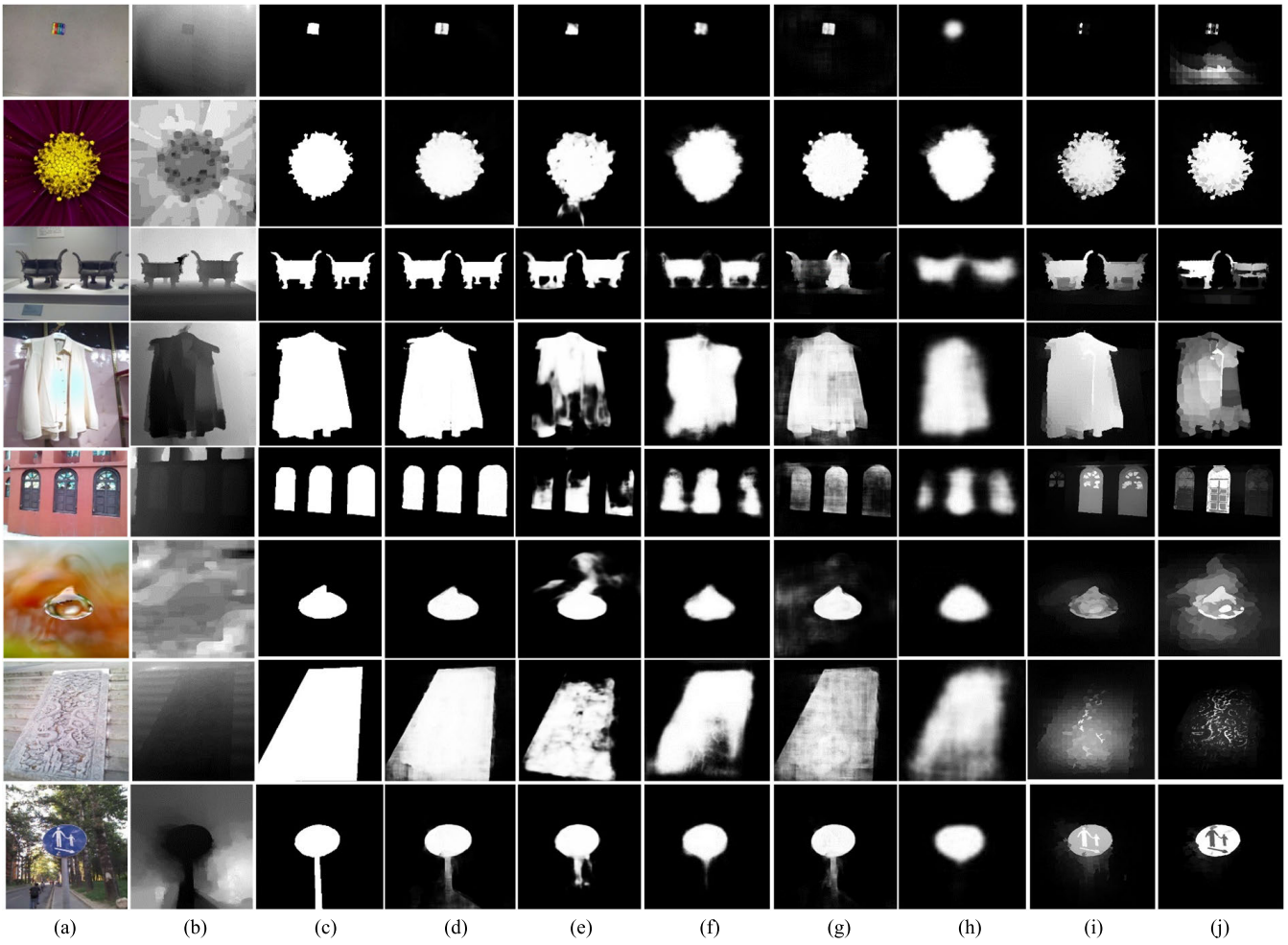


FIGURE 5. Visualization results by using various saliency detection methods. (a) RGB. (b) Depth. (c) Ground truth. (d) Ours. (e)CPFP. (f) PCF. (g) PDNet. (h) CTMF. (i) DF. (j) CDCP.

fusion module has the stronger ability to capture spatial and semantic information. This structure effectively suppresses the noises of redundant local detail and significantly improves edge performance.

4) EFFECT OF HYBRID LOSS

We give the result of PGHF model training with different losses on NJUD2K datasets, please see the Table 2. We find that Binary Cross-Entropy loss ℓ_{bc} get better S-measure than Dice-score loss ℓ_{cd}, which may infer ℓ_{bc} is more sensitive to salient structure than ℓ_{cd}. While for another three metrics (meanF, maxF and MAE), the model training with ℓ_{cd} shows the better performance than ℓ_{bc}, which may owing to ℓ_{cd} pay more attention on the unbalanced proportion of pixels in the salient object region and background area. The hybrid loss achieved further improvement on four metrics as expected.

E. COMPARISON WITH THE STATE-OF-THE-ART MENTHODS

We compare our proposed method with 6 state-of-the-art methods, including CDCP [34], DF [35], CTMF [36], PDNet [21], PCF [37] and CPFP [38]. Among them, CDCP is traditional methods which is based on handcrafted features. DF and CTMF are Early-fusion and Late-fusion models, respectively. PDNet is the first pre-trained network for RGBD SOD which learn from massive RGB datasets. PCF and CPFP are two well-known hierarchical fusion models. For fair comparison, the division of datasets and partial comparative experimental data refer to CPFP. Table 3 reports quantitative results of six methods and our proposed method on five popular RGBD datasets. We can see that our method brings better performance than other state-of-the-art methods.

To further verify the advantages of our proposed method, we also summarize several challenging situations in salient object detection, visual comparisons are showed in Fig. 5, including small object (the 1st row), multiple object (the 5th row), transparent object (the 6th row), salient object with complex edge (the 2nd and 3rd rows), salient object with center bias (the 2nd row), uneven brightness (the 4th row), low contrast (the 7th row) and complex scene (the 8th row). For small salient object, our method can depict clear boundaries and obtain almost the same shape of the small object. When the objects have center bias or complex edge, our proposed

method also produce clearer boundaries than the other methods. In the situation of transparent objects and low contrast, most of the above-mentioned approaches fail to present salient object well, but we can obtain the accurate salient regions and coherent edge. We also sample some images that contain multiple salient objects, our proposed method can completely distinguish all salient objects. Although the CPFP and PCF methods treat RGB and depth cues differently for multi-scale fusion, the fine details of the salient objects are lost severely due to the limitation of labeled RGBD datasets. Taking all the circumstances into account, our model not only locate the salient object from a global view, but also obtains fine local detail information of the object boundaries.

V. CONCLUSION

In this paper, we propose a new pre-trained guide hierarchical fusion network for RGBD salient object detection. We utilize representations learned from a large-scale RGB dataset to boost the learning ability of the model under insufficient labeled RGBD training data. Also, we design a hierarchical feature fusion module to fuse the complementary features layer by layer. Within this framework, sufficient multi-modal complementary information interactions greatly improve the network performance for salient object detect especially for contour localization. Comprehensive evaluations on five publically available benchmark datasets and comparisons with other state-of-the-arts demonstrated the effectiveness of the proposed approach.

REFERENCES

[1] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, May 2014.

[2] H. Zhang, X. Cao, and R. Wang, "Audio visual attribute discovery for fine-grained object recognition," in *Proc. AAAI*, Apr. 2018, pp. 7542–7549.

[3] C.-H. Chang, C.-K. Liang, and Y.-Y. Chuang, "Content-aware display adaptation and interactive editing for stereoscopic images," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 589–601, Aug. 2011.

[4] C. Zhu, H. Kan, and L. Ge, "An innovative saliency guided ROI selection model for panoramic images compression," in *Proc. IEEE DCC*, Mar. 2018, p. 436.

[5] S. Han and N. Vasconcelos, "Image compression using object-based regions of interest," in *Proc. IEEE ICIP*, Nov. 2007, pp. 3097–3100.

[6] J. He, T. Lin, J. Feng, and S. Chang, "Mobile product search with bag of hash bits," in *Proc. IEEE CVPR*, Nov. 2011, pp. 839–840.

[7] S.-M. Hu, T. Chen, K. Xu, M.-M. Cheng, and R. R. Martin, "Internet visual media processing: A survey with graphics and vision applications," *Vis. Comput.*, vol. 29, no. 5, pp. 393–405, May 2013.

[8] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultimediaMag.*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[9] C. Zhu and G. Li, "A multilayer backpropagation saliency detection algorithm and its applications," *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 25181–25197, Oct. 2018.

[10] H. Du, Z. Liu, H. Song, L. Mei, and Z. Xu, "Improving RGBD saliency detection using progressive region classification and saliency fusion," *IEEE Access*, vol. 4, pp. 8987–8994, 2016.

[11] F. Liu, X. Li, Y. Geng, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE CVPR*, Jun. 2012, pp. 454–461.

[12] R. Ju, L. Ge, W. Geng, and T. Ren, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE ICIP*, Oct. 2014, pp. 1115–1119.

[13] D. Feng, N. Barnes, S. You, and C. Mccarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. IEEE CVPR*, Jun. 2016, pp. 2343–2350.

[14] Y. Fang, J. Wang, Y. Yuan, J. Lei, W. Lin, and P. L. Callet, "Saliency-based stereoscopic image retargeting," *Inf. Sci.*, vol. 372, pp. 347–358, Dec. 2016.

[15] J. Guo, T. Ren, J. Bei, and Y. Zhu, "Salient object detection in RGB-D image based on saliency fusion and propagation," in *Proc. 7th Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2015, pp. 1–5.

[16] H. Song, Z. Liu, Y. Xie, L. Wu, and M. Huang, "RGBD co-saliency detection via bagging-based clustering," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1722–1726, Dec. 2016.

[17] M. García, E. Potapova, T. Werner, M. Zillich, M. Vincze, and S. Frintrop, "Saliency-based object discovery on RGB-D data with a late-fusion approach," in *Proc. IEEE ICRA*, May 2015, pp. 1050–4729.

[18] 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.*, to be published.

[19] Z. Li, C. Lang, S. Feng, and T. Wang, "Saliency ranker: A new salient object detection method," *J. Vis. Commun. Image Represent.*, vol. 50, pp. 16–26, Jan. 2018.

[20] H. Fu, D. Xu, and S. Lin, "Object-based multiple foreground segmentation in RGBD video," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1418–1427, Mar. 2017.

[21] C. Zhu, X. Cai, K. Huang, and T. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *Proc. IEEE ICME*, Jul. 2019, pp. 199–204.

[22] 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.

[23] R. Achantay, S. Hemamiz, F. Estrada, and S. Süsstrunky, "Frequency-tuned salient region detection," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1597–1604.

[24] T. Cuong and D. Chandler, "An algorithm for detecting multiple salient objects in images via adaptive feature selection," in *Proc. IEEE ICIP*, Sep. 2012, pp. 657–660.

[25] MM. Cheng, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[26] Y. Niu, C. Su, and W. Guo, "Salient object segmentation based on superpixel and background connectivity prior," *IEEE Access*, vol. 6, pp. 56170–56183, 2018.

[27] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE CVPR*, Oct. 2015, pp. 5455–5463.

[28] L. Yi, "Feature pyramid networks for object detection," in *Proc. IEEE CVPR*, Jul. 2017, pp. 2117–2125.

[29] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE CVPR*, Jun. 2016, pp. 678–686.

[30] Y. Zhuge, G. Yang, P. Zhang, and H. Lu, "Boundary-guided feature aggregation network for salient object detection," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1800–1804, Dec. 2018.

[31] A. Ciptadi, T. Hermans, and J. Rehg, "An in depth view of saliency," in *Proc. BMVC*, Sep. 2013, pp. 112.1–112.11.

[32] X, Fan, Z. Liu, and G. Sun, "Salient region detection for stereoscopic images," in *Proc. Digital Signal Process.*, 2014, pp. 454–458.

[33] J. Ren, "Exploiting global priors for RGB-D saliency detection," in *Proc. IEEE CVPR*, Jun. 2015, pp. 25–32.

[34] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Proc. IEEE Comput. Vis.*, Oct. 2017, pp. 1509–1515.

[35] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.

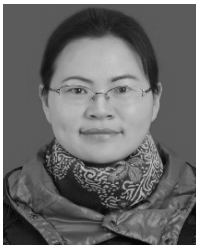
[36] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.

[37] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE CVPR*, Jun. 2018, pp. 3051–3060.

[38] J. Zhao, Y. Cao, D. Fan, and M. M. Cheng, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proc. IEEE CVPR*, Jun. 2019, pp. 3927–3936.

[39] F. Xiao, W. Deng, L. Peng, C. Cao, K. Hu, and X.-P. Gao, "MSDNN: Multi-scale deep neural network for salient object detection," *IET Image Process.*, vol. 12, no. 11, pp. 2036–2041, Nov. 2018.

[40] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.
[41] L. Fidon, W. Li, and L. Garcia-Peraza-Herrera, "Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2017, pp. 64–76.
[42] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, vol. 2014, pp. 92–109.
[43] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1605–1616, Aug. 2017.
[44] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. ICMCS*, Jul. 2014, p. 23.
[45] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Comput. Sci., Tech. Rep. CSTR, Feb. 2005, vol. 2, no. 11, pp. 1–11.
[46] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE ICCV*, Oct. 2017, pp. 4558–4567.
[47] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1597–1604.
[48] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE CVPR*, Jun. 2012, pp. 733–740.
[49] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comp. Vis. Media*, vol. 5, no. 2, pp. 117–150, Jun. 2019.



FEN XIAO was born in 1981. She received the B.S. degree in computer science and the Ph.D. degree in computational mathematics from Xiangtan University, China, in 2002 and 2008, respectively. She was a Visiting Scientist with the Department of Energy's Pacific Northwest National Laboratory (PNNL), WA, USA, from 2015 to 2016. She is currently a Professor with the College of Computer Science, Xiangtan University. She has authored more than 20 journal articles and conference papers. Her research interests include image processing, image captioning, and neural networks.

BIN LI was born in 1995. He received the B.S. degree from the College of Physics and Electronic Engineering, Hainan Normal University, China, in 2017. He is currently pursuing the M.S. degree in information and communication engineering from Xiangtan University, China. His research interests are saliency detection, image semantic recognition, and deep neural networks.

YIMU PENG was born in 1994. He received the B.S. degree in information and communication engineering from Xiangtan University, China, in 2017, where he is currently pursuing the M.S. degree in information and communication engineering. His research interests are saliency detection, image semantic recognition, and deep neural networks.

CHUNHONG CAO received the B.S. degree in applied mathematics and software from Central South University, Changsha, China, in 1999, and the M.S. and Ph.D. degrees from Xiangtan University, Xiangtan, China, in 2005 and 2017, respectively. She is currently an Associate Professor with the College of Computer Science, Xiangtan University, Xiangtan, China, and also a Visiting Scholar with the Department of Computer Science, University of Georgia, Athens, GA, USA. Her current research interests include hyperspectral image processing, wavelet analysis theory, and compressive sensing.

KAI HU received the B.S. degree in computer science and the Ph.D. degree in computational mathematics from Xiangtan University, Hunan, China, in 2007 and 2013, respectively. He was a Visiting Scholar with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2016 to 2017. He is currently an Associate Professor with the College of Computer Science, Xiangtan University. His current research interests include machine learning, pattern recognition, and medical image processing.

XIEPING GAO (Member, IEEE) was born in 1965. He received the B.S. and M.S. degrees from Xiangtan University, China, in 1985 and 1988, respectively, and the Ph.D. degree from Hunan University, China, in 2003. He was a Visiting Scholar with the National Key Laboratory of Intelligent Technology and Systems, Tsinghua University, China, from 1995 to 1996, and also with the School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, from 2002 to 2003. He is currently a Professor with the College of Computer Science, Xiangtan University, China. He has authored and coauthored more than 80 journal articles, conference papers, and book chapters. His current research interests are in the areas of wavelets analysis, neural networks, evolution computation, and image processing. He is a Regular Reviewer for several journals and he has been a member of the technical committees of several scientific conferences.