ASIF-Net: Attention Steered Interweave Fusion Network for RGB-D Salient Object Detection

Chongyi Li, Runmin Cong, *Member, IEEE*, Sam Kwong, *Fellow, IEEE*, Junhui Hou, *Member, IEEE*, Huazhu Fu, *Senior Member, IEEE*, Guopu Zhu, *Senior Member, IEEE*, Dingwen Zhang, and Qingming Huang, *Fellow, IEEE*

***Abstract*—Salient object detection from RGB-D images is an important yet challenging vision task, which aims at detecting the most distinctive objects in a scene by combining color information and depth constraints. Unlike prior fusion manners, we propose an attention steered interweave fusion network (ASIF-Net) to detect salient objects, which progressively integrates cross-modal and cross-level complementarity from the RGB image and corresponding depth map via steering of an attention mechanism. Specifically, the complementary features from RGB-D images are jointly extracted and hierarchically fused in a dense and interweaved manner. Such a manner breaks down the barriers of inconsistency existing in the cross-modal data and also sufficiently captures the complementarity. Meanwhile, an attention mechanism is introduced to locate the potential salient regions in an attention-weighted fashion, which advances in highlighting the salient objects and suppressing the cluttered background regions. Instead of focusing only on pixelwise saliency, we also ensure that the detected salient objects have the objectness characteristics (e.g., complete structure and sharp boundary) by incorporating the adversarial learning that provides a global semantic constraint for RGB-D salient object detection. Quantitative and qualitative experiments demonstrate that the proposed method performs favorably against 17 state-of-the-art saliency detectors on four publicly available RGB-D salient object detection datasets. The code and results of our method are available at https://github.com/Li-Chongyi/ASIF-Net.**

***Index Terms*—Adversarial learning, depth cue, interweave fusion, residual attention, RGB-D images, saliency detection.**

Manuscript received June 20, 2019; revised November 14, 2019; accepted January 18, 2020. This work was supported in part by the Dr. Cong's Project of the Fundamental Research Funds for the Central Universities under Grant 2019RC039, in part by the National Natural Science Foundation of China under Grant 61771334, Grant 61871342, Grant 61872350, Grant 61672443, Grant 61931008, Grant 61836002, and Grant U1636214, in part by the Hong Kong Research Grants Council General Research Funds under Grant 9042038 (CityU 11205314) and Grant 9042322 (CityU 11200116), in part by the Hong Kong Research Grants Council Early Career Schemes under Grant 9048123 (CityU 21211518), and in part by the China Postdoctoral Support Scheme for Innovative Talents under Grant BX20180236. This article was recommended by Associate Editor H. Lu. *(Chongyi Li and Runmin Cong contributed equally to this work.) (Corresponding author: Runmin Cong.)*

Chongyi Li is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: lichongyi25@gmail.com).

Runmin Cong is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: rmcong@bjtu.edu.cn).

Sam Kwong and Junhui Hou are with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 51800, China (e-mail: cssamk@cityu.edu.hk; jh.hou@cityu.edu.hk).

Huazhu Fu is with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE (e-mail: hzfu@ieee.org).

Guopu Zhu is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: gp.zhu@siat.ac.cn).

Dingwen Zhang is with the School of Mechano-electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: zhangdingwen2006yyy@gmail.com).

Qingming Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: qmhuang@ucas.ac.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2020.2969255

I. Introduction

SALIENT object detection aims at identifying the most attractive parts in an image, which has been widely used in many computer vision and machine intelligence tasks [1]–[3], such as segmentation [4], enhancement [5], foreground annotation [6], thumbnail creation [7], and quality assessment [8]. With different input data, the saliency detection task can be roughly divided into image saliency detection for an individual image [9]–[21], co-saliency detection for an image group, including multiple-related images [22]–[31], video saliency detection for the video sequences [32]–[40], and saliency detection for light field [41]–[43]. In fact, humans can perceive the depth information of the scene through the binocular vision system. In recent years, depth information has become increasingly popular due to the rapid development of affordable and portable consumer depth cameras [44], [45], which can provide many useful cues, such as shape and boundary. Introducing the depth constraint to saliency detection can benefit addressing some challenging situations, for example, the salient object and background sharing the similar appearance, the complex and cluttered background, etc. However, there are two main issues that need to be addressed in RGB-D salient object detection: 1) how to effectively and sufficiently integrate the cross-modal complementarity from the RGB image and corresponding depth map and 2) how to accurately locate the salient object with complete structure and sharp boundary. To address these two issues, we propose an attention steered interweave fusion network (ASIF-Net) for RGB-D salient object detection.

To accurately formulate the complementarity from RGB-D images and effectively fuse the cross-modal information, several fusion manners have been proposed, such as input fusion [46], early fusion [16], late fusion [47], etc. The input fusion manner concatenates the RGB image and corresponding

depth map as the input of networks. The early or late fusion manner first extracts the RGB and depth features separately. Then, early fusion manner fuses the cross-modal features in the shallow layers of networks, while the late fusion manner is to fuse the RGB and depth features in the deep layers of networks. It has been widely demonstrated that the features in deep layers carry more semantic information while the features in shallow layers have more detailed information [48], [49]. Moreover, both the high-level semantic information and low-level detailed information are essential for improving the performance of salient object detection. Thus, instead of performing the previous fusion manners which are insufficient and brute forcing, we propose an interweave fusion scheme, which mutually guides the joint extraction of the RGB-D features and densely integrates cross-modal complementarity through the shallow and deep layers of our network. In this way, the proposed network automatically and sufficiently grabs the potentially useful information and reduces the interference induced by the inconsistency of cross-modal data.

In addition, most of the salient object detection methods treat all regions in the extracted features equally and ignore the fact that different regions should have different contributions to the final detection. As a result, these methods usually carry redundant features and are easily affected by the cluttered background. To address this issue, we introduce an in-block residual attention module (In-b RAM) to weigh the importance of different regions in the jointly extracted features. The proposed In-b RAM treats the attention weight as identical mapping (residual attention) in a deep side output supervision way, which is different from the widely used self-attention mechanism which calculates response at a position as a weighted sum of the features at all positions [50] and contextual attention mechanism where each attention weight corresponds to the contextual relevance at each context location [18]. In addition, instead of upsampling the side outputs to the same resolution as the layers in the shallower block by using the bilinear interpolation like in [51], we generate the side outputs by the joint features in the same block (in-block), which leads to the sharp and clear salient objects in the final result. This is because, we found that the bilinear interpolation tends to blur the side output, especially for its edges and boundaries.

The existing deep-learning-based RGB-D salient object detection methods usually minimize the cross-entropy loss functions (e.g., binary cross-entropy (BCE) loss and balanced cross-entropy loss) to optimize the deep networks. However, these loss functions do not go beyond the limitations of pixelwise detection, which may induce incomplete and unclear salient objects. Thus, instead of focusing only on pixelwise saliency, we also ensure the saliency map is indistinguishable from the ground truth to a given discriminator by incorporating the adversarial learning into our framework.

In summary, our main contributions are listed as follows.

1) We propose an ASIF-Net architecture, which progressively and interactively captures complementarity from RGB-D images via the interweave fusion and weighs the saliency regions by the steering of the deeply supervised attention mechanism.
2) We design an In-b RAM to avoid the blur induced by upsampling interpolation and introduce a global semantic constraint to our network by using adversarial learning. By doing these, our method can effectively suppress the cluttered background regions and obtain the complete structure and sharp boundary of salient objects.
3) Without the use of any preprocessing and post-processing strategies, the proposed network is trained from scratch and achieves the competitive performance against 17 state-of-the-art saliency detectors on four benchmark datasets.

II. Related Work

In this section, we briefly review the related works, including RGB image saliency detection and RGB-D image saliency detection.

A. RGB Image Saliency Detection

The last decades hae witnessed the encouraging development and improvement of saliency detection for the RGB image. Numerous bottom-up models [9]–[14] and top-down models [15]–[21] have been presented in recent years. The background prior has been introduced to achieve saliency detection in [10]. Zhou *et al.* [11] proposed a bottom-up method for detecting salient regions in images by propagating the saliency and background seed vectors on a new two-layer sparse graph. Huang *et al.* [12] formulated the saliency detection problem as a multiple instances learning (MIL) task by taking the proposals as the bags of instances of MIL, where the instances are the superpixels contained in the proposals. Peng *et al.* [13] formulated saliency detection as a structured matrix decomposition problem guided by high-level priors. Yuan *et al.* [14] achieved the saliency detection method with reversion correction and regularized random walk ranking. Recently, deep learning has demonstrated superior performance in saliency detection. Hou *et al.* [16] introduced short connections into the skip-layer structures within the holistically nested edge detector architecture to achieve saliency detection. Hu *et al.* [17] introduced the recurrently aggregated deep features into a fully convolutional network to achieve saliency detection by fully exploiting the complementary saliency information captured in different layers. Deng *et al.* [19] proposed a recurrent residual refinement network for saliency detection, where some residual refinement blocks are designed to recurrently learn the difference between the coarse saliency map and ground truth by alternatively harnessing the low-level and high-level features. Considering the boundary quality in saliency detection, Qin *et al.* [20] proposed a predict-refine architecture, which consists of a densely supervised encoder–decoder network and a residual refinement module. Similarly, for preserving salient object boundaries well, Zhao *et al.* [21] integrated the complementary features between salient edge information and salient object information in a single network to achieve salient object detection.

B. RGB-D Image Saliency Detection

From a depth map, one can capture many useful attributes to assist in identifying the salient object from the complex

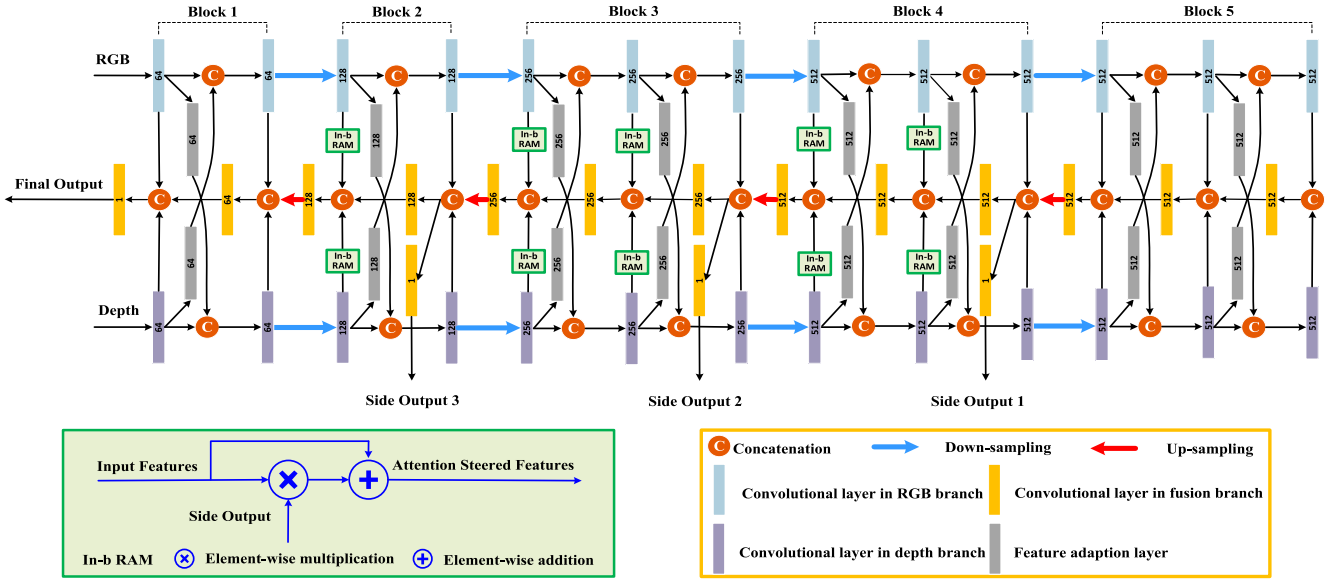

Fig. 1. Overview architecture of the proposed network. Our network consists of three main branches: 1) RGB branch to extract the RGB features and meanwhile, fuse the guidance features from the same level in the depth branch; 2) depth branch to extract the depth features and meanwhile, fuse the guidance features from the same level in the RGB branch; 3) fusion branch to gradually integrate the complementary RGB-D features and weight the importance of different regions in the RGB-D features by the deeply supervised side outputs. In-b RAM which uses the side output to weight the RGB-D features in the same block. The details of the In-b RAM are shown in the lower left corner. Downsampling is implemented by 2× max pooling while upsampling is implemented by transposed convolution with kernels of size 3 × 3 and stride 2. The convolutional layer has the same kernels of size 3 × 3 and stride 2 and the number in each convolutional layer indicates the number of output feature maps. All convolutional layers are followed by the rectified linear unit (ReLU) activation function [52], except for the convolutional layers before outputs (using sigmoid activation function). Best viewed with zoom in on a digital display.

background, such as the shape attribute, boundary cue, and surface normal. The previous works mainly focus on extracting handcrafted features to represent the salient region, where the depth cue is utilized as an explicit supplement to color feature [46], [53]–[55] or an implicit expression through some designed measures [56]–[59]. Peng *et al.* [46] proposed a multistage saliency model for RGB-D images by combining the low-level feature contrast, mid-level region grouping, and high-level prior enhancement. Song *et al.* [54] presented a salient object detection framework for RGB-D images via the multiscale discriminative saliency fusion, where the depth contrast is treated as a common depth property. Ju *et al.* [56] designed an anisotropic center-surround difference (ACSD) measure to calculate the depth-aware saliency map. Feng *et al.* [57] proposed a local background enclosure (LBE) measure to directly capture salient structure from the depth map. Cong *et al.* [58] considered the quality of depth map and proposed a depth confidence measure to reduce the negative influence of poor depth map in saliency detection.

Recently, with the fast development of deep-learning strategy in diverse visual tasks [60]–[62], deep learning has been successfully applied to RGB-D saliency detection and achieved admirable performance. Qu *et al.* [63] designed a convolutional neural network (CNN) to learn the interaction between the low-level raw saliency features and saliency result for RGB-D saliency detection, and a superpixel-based Laplacian propagation model was introduced to improve the spatial consistency of saliency map. Han *et al.* [47] transferred the structure of the RGB deep network to the depth view and fused the deep representations of both views to generate the final saliency map. Chen and Li [64] proposed a progressively complementarity-aware fusion network for RGB-D salient object detection, where the cross-modal residual functions and complementarity-aware supervision are introduced. Chen *et al.* [65] presented a multiscale multipath fusion network for RGB-D saliency detection, which advances the traditional two-stream fusion architecture. Chen and Li [66] proposed a three-stream attention-aware network for RGB-D salient object detection, where the cross-modal distillation stream focuses on augmenting the RGB-D representation in the bottom-up path, and the channel-wise attention mechanism aims at adaptively selecting complementary feature maps in the top-down inference path. Zhao *et al.* [67] utilized contrast prior into CNNs-based architecture to enhance the depth information, and the enhanced depth cues are further integrated with RGB features for salient object detection through a novel fluid pyramid integration module. Piao *et al.* [68] designed a depth-induced multiscale recurrent attention network for salient object detection, which consists of a depth refinement block and a recurrent attention module. Chen and Li [66] proposed a three-stream attention-aware network by integrating the cross-modal distillation stream in the bottom-up path and the channel-wise attention in the top-down inference path.

III. Methodology

A. Pipeline

Fig. 1 illustrates the overview architecture of the proposed network which consists of three main branches, that is, an RGB branch, a depth branch, and a fusion branch. Specifically, the RGB branch extracts multilevel features from an input

RGB image and meanwhile, integrates these features with the depth features from the same level depth branch, *vice versa* for the feature extraction and integration process in the depth branch. Before the feature integration in the RGB/depth branch, the feature adaption layers are used to transform the RGB/depth features for adapting the depth/RGB branch, which relieves the inconsistency between different modalities. In the fusion branch, we progressively integrate the complementary RGB-D features from the same level in a deep-to-shallow manner (from block 5 to block 1) and employ deep supervision to produce side outputs in a coarse-to-fine fashion (from side output 1 to side output 3). With these side outputs, an In-b RAM is designed to weigh the importance of different regions in the RGB-D features which have the same resolution to the corresponding side output. Finally, the network generates the saliency map by using these discriminatively complementary feature representations. Next, we will introduce the interweave fusion, In-b RAM, and loss function in detail.

B. Interweave Fusion

Considering the inconsistency of cross-modal data, our interweave fusion performs dense interaction and hierarchical fusion between RGB-D features. Instead of independently extracting the RGB-D features and then integrating them which ignore the connections of paired RGB-D features, an interaction guidance way between RGB-D features can facilitate the integration of cross-modal complementarity and remit the inconsistency existed in different modalities. Besides, it also has the potential for alleviating the negative influence of the degraded depth maps induced by the imaging environment and devices.

Specifically, both the RGB and depth branches share the same backbone which includes five blocks. The downsampling operation is the boundary of each block. In each block, the features have the same dimension. Besides, except the trunk, each branch contains eight feature adaption layers (i.e., extra convolutional layers) denoted in gray in Fig. 1. These feature adaption layers transform the current RGB (resp. depth) features, which are used to guide the extraction of deeper depth (resp. RGB) features. Such an interaction guidance fusion is capable of reducing the inconsistency and facilitating the integration of complementarity from cross-modal features. Also, it provides robust and compact features, which can be expressed as

$$f_{rgb}^{c} = \left\{ f_{rgb}, g_d(f_d) \right\} \tag{1}$$

where f_{rgb}^{c} denotes the concatenated features of the current RGB features f_{rgb} and the guidance features $g_d(f_d)$ from depth branch. g_d is the feature adaption layer in the depth branch, which is implemented by a convolutional layer with the kernels of size 3×3 and stride 2. f_d denotes the corresponding depth features in the same level. For f_{rgb} and f_d, the subscripts rgb and d mean the features are from RGB branch and depth branch, respectively. $\{\cdot,\cdot\}$ represents the concatenation operation which is the stack of feature maps with the same shape except for the concatenation axis. The interaction guidance fusion in the depth branch has the similar expression. With the hierarchical cross-modal features, a dense fusion fashion, occurring among the RGB features and depth feature in the same level and the former fused features, is proposed to extract the discriminative representations. Such a dense fusion manner sufficiently integrates the complementary features from cross-modal data and forces the fused features in the deeper layer toward the final output in a coarse-to-fine manner fashion. To better understand the pipeline of our method, we present the examples of our network outputs in Fig. 2. It is clear that the outputs of our network from side output 1 to final output gradually become complete and sharp with the deep-to-shallow attention steered interweave fusion and global semantic constraint. For example, in the first image, all side outputs can detect the coracoid region, although it shares the similar color to the background. In particular, the final output yields a complete structure and sharp boundary.

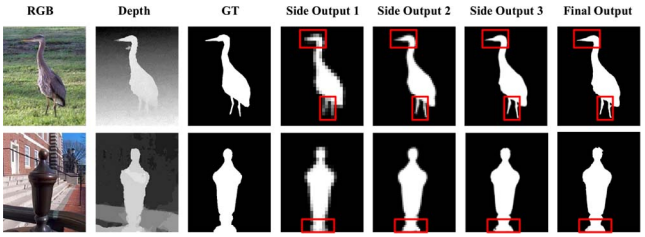

Fig. 2. Examples of our network outputs. From left to right are the RGB images, corresponding depth maps, ground truth, our network side output 1, side output 2, side output 3, and final output. For comparisons, the side outputs are resized to the same size as the final output.

C. In-Block Residual Attention Module

Based on the fact that different regions may have different contributions to the saliency prediction, we introduce an In-b RAM to the process of cross-modal fusion, where the side outputs are treated as the feature selectors to weight the features of the shallower layers in the same block. By doing this, the attention mechanism is helpful for better highlighting the salient objects and suppressing the cluttered background.

Specifically, for each In-b RAM, the inputs are the features from the RGB branch or depth branch. First, the inputs multiply the corresponding attention map, which is the side output in the same block, to obtain the weighted inputs. Then, the inputs add the weighted inputs to obtain the outputs of In-b RAM (identical mapping fashion). The potential reasons for considering the attention module as an identical mapping are: 1) a deep network should not produce worse performance than its counterpart without the attention module; 2) directly feeding the weighted features to the following layers may lead to gradient vanishing in the backpropagation since the values of the input features may become very small or even vanish after multiplying the attention map; and 3) the identical mapping can keep good properties of original features. The details of the In-b RAM are shown in the lower left corner of Fig. 1, which can be expressed as

$$f_{\text{inbram}} = f \oplus f \otimes W_{sp} \tag{2}$$

where f_{inbram} denotes the features after the In-b RAM, f is a set of input features, W_{sp} indicates the side output in the same block as the input features, and $\oplus$ and $\otimes$ represent the

elementwise addition and elementwise multiplication, respectively. Note that we did not use the In-b RAM in the deepest block because the resolution of features in this block is too low to provide sufficient information. The advantage of our in-block RAM will be further demonstrated in the ablation studies.

D. Loss Function

BCE Loss: Given the input RGB-D images $X = \{x_i, i = 1, \ldots, T\}$, the corresponding binary saliency map is denoted as $Y = \{y_i, i = 1, \ldots, T\}$, where x_i and y_i stand for the pixel values, and T represents the number of pixels. To calculate the loss of the side outputs, we successively downsample Y to obtain scaled versions of the original binary saliency map corresponding to the side outputs. Following the previous works, we employ the standard BCE loss on both the final and side outputs, which can be expressed as

$$L_f^{bce} = -\sum_{i \in Y_+} \log P(y_i = 1|X; \mathbf{\Phi}) - \sum_{i \in Y_-} \log P(y_i = 0|X; \mathbf{\Phi}) \quad (3)$$

where $\mathbf{\Phi}$ are the parameters of the proposed network, which need to be learned. Y_+ and Y_- denote the positive and negative labeled pixel sets in the saliency map Y, and $P(\cdot)$ is the saliency confidence score. Similar to the L_f^{bce}, the losses of the side outputs 1–3 are denoted as L_{s1}^{bce}, L_{s2}^{bce}, and L_{s3}^{bce}, respectively.

Adversarial Loss: Most of the learning-based salient object detection methods only use the pixelwise loss functions, which ignore the global information and objectness attributes of the salient objects. In other words, we should build the relations of salient pixels and highlight the entire object from a global perspective. Therefore, we incorporate the adversarial learning into our framework to provide a global semantic constraint that constrains the generated salient objects having the complete structure and sharp boundaries but does not increase the network complexity.

Specifically, to leverage the context information, we concatenate the final output with the RGB-D images that the current output corresponds to as the inputs. Then, the inputs are fed to a discriminator to learn a joint distribution. To encourage the generated saliency map to be indistinguishable from the ground truth, we solve the following optimization problem:

$$\min_G \max_D \mathbb{E}_I\big[\log(1 - D(I, G(I)))\big] + \mathbb{E}_{I,J}\big[\log D(I, J)\big] \quad (4)$$

where G is the generator (i.e., our ASIF-Net) which tries to generate "fake" images to fool the discriminator, D is the discriminator which tries to distinguish fake images from the real images, I are the inputs (i.e., the combination of RGB-D images and generated saliency map), and J are the corresponding ground truth (i.e., the combination of RGB-D images and real saliency map). To train the generative network, we directly minimize the following loss function:

$$L_f^{adv} = \log(1 - D(I, G(I))). \quad (5)$$

Final Loss: The final loss function for training our network is the linear aggregation of the BCE and adversarial losses

$$L_f = \lambda L_f^{adv} + L_f^{bce} + L_{s1}^{bce} + L_{s2}^{bce} + L_{s3}^{bce} \quad (6)$$

where λ is the weight of L_f^{adv}, which is picked empirically based on preliminary experiments on the training data. In our network, the side outputs are relevant, which are used to weight the features in the same block, and then, the weighted features are forwarded to the next block for predicting subsequent side output. Thus, we give these side outputs the same importance in our network. Note that we only apply the adversarial loss to constrain the final output because it is hard to achieve the convergence when putting the adversarial loss on each side output.

IV. Experiments

In this section, we first describe the benchmark datasets and evaluation metrics, and then illustrate the training strategies and implementation details of the proposed network. In addition, we compare the proposed method with state-of-the-art salient object detection methods to demonstrate its advantages. Finally, a series of ablation studies are conducted to verify the role of each component of our proposed network.

A. Benchmark Datasets and Evaluation Metrics

Benchmark Datasets: To evaluate the performance of the proposed method, we conduct comprehensive experiments on four datasets, that is, NJUD [56], NLPR [46], STEREO [53], and LFSD [41].

1) The *NJUD* dataset contains 2003 RGB images and corresponding depth maps with diverse objects and complex scenarios, as well as the pixelwise ground truth for saliency detection. The depth maps are estimated from the stereo images.
2) The *NLPR* dataset contains 1000 RGB-D images with pixelwise ground truth, where the depth maps are captured by Microsoft Kinect under different illumination conditions and acquisition scenes. Moreover, multiple salient objects may exist in an image of this dataset.
3) The *STEREO* dataset collects 797 paired RGB-D images and annotates the corresponding ground truth saliency masks, where the depth maps are estimated from the stereo images.
4) The *LFSD* dataset is constructed for light field saliency detection, which contains 100 all-focus RGB images, the corresponding depth maps, and the pixelwise ground truth masks. The depth map was captured by the Lytro light field camera.

Evaluation Metrics: For quantitative evaluations, the Precision–Recall (P–R) curve, F-measure, MAE score, and S-measure are employed. Thresholding the saliency map at a series of fixed integers from 0 to 255, we can obtain some binary saliency masks. Then, the precision and recall scores are calculated by comparing the binary mask with the ground truth. Thus, the P–R curve is drawn under different combination of precision and recall scores. F-measure

as a comprehensive measurement is defined as the weighted harmonic mean of precision and recall [69]

$$F_\beta = \frac{(1+\beta^2)\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{7}$$

where β^2 is set to 0.3 for emphasizing the precision as suggested in [69].

MAE score [1] calculates the difference between the continuous saliency map S and ground truth G, which indicates how similar a saliency map is compared to the ground truth. The MAE score can be calculated as

$$\text{MAE} = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} |S(i,j) - G(i,j)| \tag{8}$$

where w and h correspond to the width and height of the image, respectively.

S-measure [70] evaluates the structural similarity between the saliency map and ground truth as

$$S_m = \alpha \times S_o + (1-\alpha) \times S_r \tag{9}$$

where α is set to 0.5 for assigning equal contribution to both region (S_r) and object (S_o) similarity.

B. Training and Implementation Details

Training: For a fair comparison, we adopt the same training, validation, and testing sets as in [47] and [64]. The training set includes 650 samples from the NLPR dataset and 1400 samples from the NJUD dataset. The training samples are resized to 224 × 224. Moreover, we augment the training data with flipping and rotation. For the validation set, 50 samples are randomly selected from the NLPR dataset and 100 samples from the NJUD dataset. The remaining samples in these four datasets are used for testing.

Implementation: We implement the proposed network with TensorFlow on a PC with an Intel i7 6700 CPU, 32-GB RAM, and an NVIDIA GeForce GTX 1080Ti GPU. During training, a batch-mode learning method with a batch size of eight is applied. The filter weights of each layer are initialized by Xavier, and the bias is initialized as a constant. We use ADAM for network optimization, and fix the learning rate to $1e^{-4}$ in the entire training procedure. We iteratively train the generator and discriminator in the optimization procedure until convergence. To be specific, we only train the generator in the first 10 epochs. After that, we iteratively update the generator and discriminator. Finally, the network is convergent after 32 epochs. report the learning curve of our network on the validation set in Fig. 3. It took around 20 h to optimize our network. The discriminator uses the same architecture and parameter settings as the discriminator in [71]. The weight λ is set to 0.01. The backbone network of our ASIF-Net is VGG-16 [72] and we train it from scratch. In addition, our network does not use any preprocessing (e.g., HHA code [73] for depth) and post-processing (e.g., CRFs [74]) strategies. The average runtime of our method is 0.063 s for an image with the size of 224×224 based on the above-mentioned configurations.

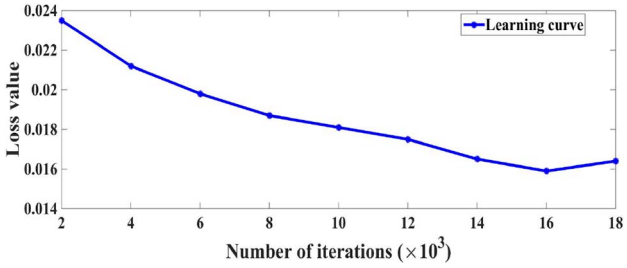

Fig. 3. Learning curve of our network on the validation set.

C. Comparison With State-of-the-Art Methods

We extensively compare the proposed method with 17 state-of-the-art methods on four datasets, including four unsupervised RGB saliency detection methods (DSG [11], MILPS [12], SMD [13], and RCRR [14]), three deep-learning-based RGB saliency detection methods (DCL [15], DSS [16], and R3Net [19]), three unsupervised RGB-D saliency detection methods (ACSD [56], DCMC [58], and MBP [55]), and seven deep-learning-based RGB-D saliency detection methods (DF [63], CTMF [47], PCFN [64], MMCI [65], CPFP [67], DMRA [68], and TANet [66]. The saliency maps of these compared methods are generated by the original codes under the default parameters or provided by the authors. Some visual examples are shown in Fig. 4. We can see the following.

1) The unsupervised RGB saliency detection methods fail to highlight the salient objects accurately due to the lack of depth constraint and supervision information. For example, the salient objects (e.g., the Athena sculpture in the seventh image, the white cat in the fourth last image, and the cake in the second last image) are not effectively detected by the SMD [13] and RCRR [14] methods. In addition, some background regions (e.g., the tower in the first image, the buildings in the distance in the fourth image, and the flame in the third last image) are wrongly highlighted as salient regions by these methods.
2) In view of the powerful learning ability of deep learning, the performance of the RGB saliency detection methods based on deep learning (i.e., DCL [15] and R3Net [19]) are obviously improved. However, for some challenging images, the salient object cannot be completely and accurately highlighted and the background regions are not effectively suppressed, such as the man wearing the striped shirt in the third image and the can in the fifth image (i.e., the multiple objects issue), the sculpture in the eighth last image, and the dog in the seventh last image (i.e., the issue of similar appearance between foreground and background), the Athena sculpture in the seventh image and the cake in the second last image (i.e., the issue of complex and cluttered background).
3) Compared with the unsupervised RGB saliency method, RGB-D saliency detection methods (e.g., ACSD [56] and DCMC [58]) achieve better performance with suppressed background by considering the depth cue. For example, the Athena sculpture in the seventh image is well detected by the ACSD [56] method and the

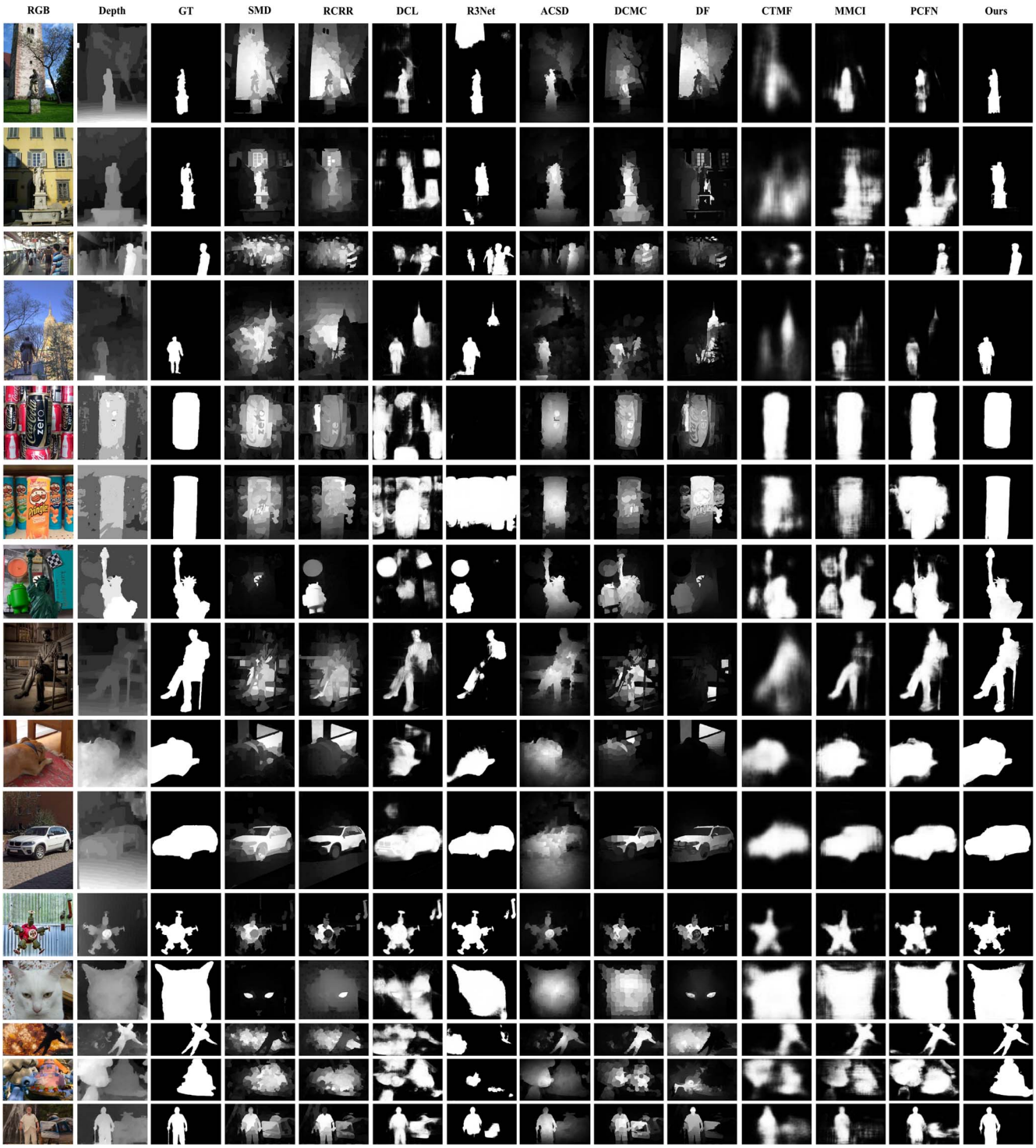

Fig. 4. Visual comparisons of different methods. From left to right are the input RGB images, the corresponding depth maps, the ground truth, and the results of SMD [13], RCRR [14], DCL [15], R3Net [19], ACSD [56], DCMC [58], DF [63], CTMF [47], MMCI [65], PCFN [64], and our ASIF-Net.

background region (i.e., the flame in the third last image) is effectively suppressed by the DCMC [58] method. However, the salient regions are not accurately located, and the background regions are not perfectly suppressed, such as the sculpture in the first image, the man in the third image, and the car in the sixth last image.

4) Combining the advantages of depth cue with the powerful ability of deep learning, some deep saliency models are proposed, such as DF [63], CTMF [47], PCFN [64], MMCI [65], CPFP [67], DMRA [68], and TANet [66]. However, these methods cannot achieve favorable and stable performance. For example, the salient regions are not highlighted with exact position, complete structure, and sharp boundaries by the DF [63] and CTMF [47] methods (e.g., the sculptures in the first four images, the dog in the seventh last image,

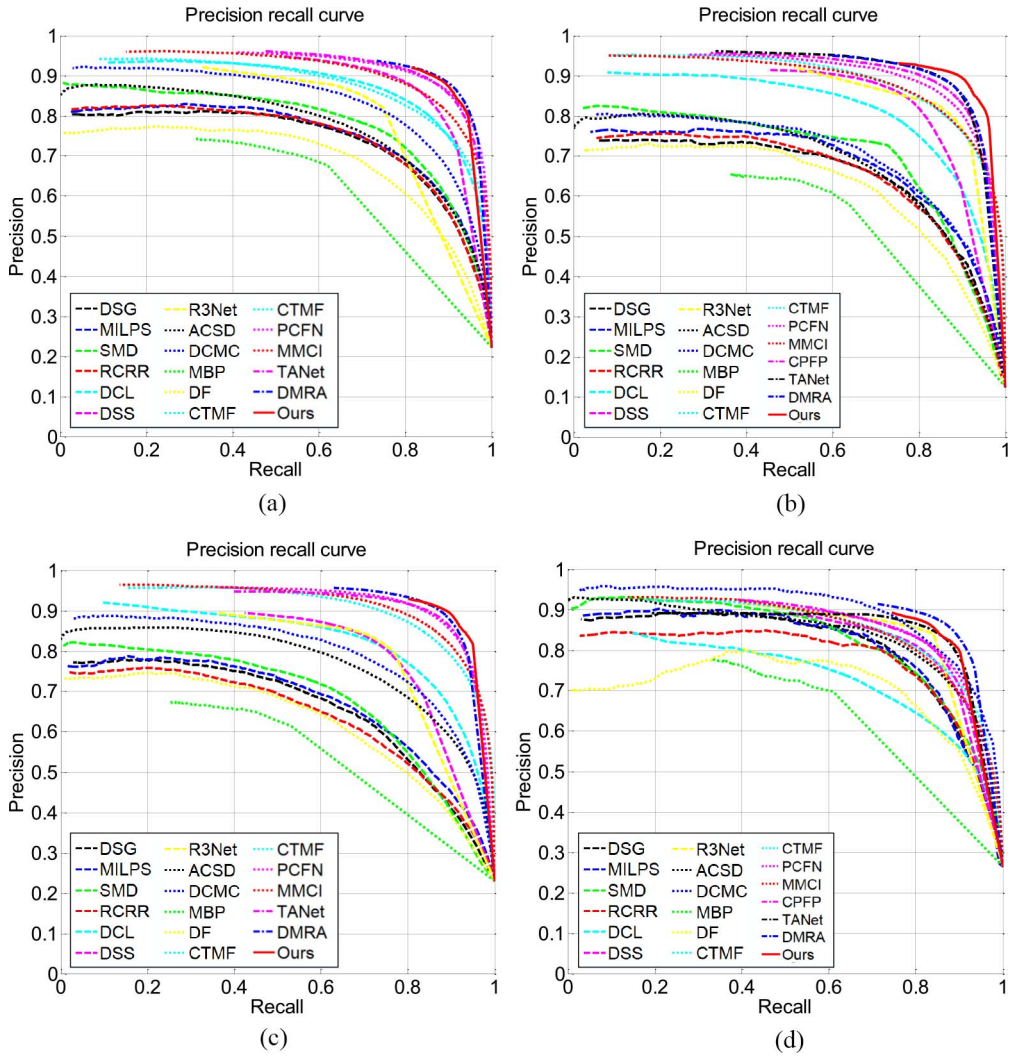

Fig. 5. P–R curves of different methods on four datasets. (a) STEREO dataset. (b) NLPR-test dataset. (c) NJUD-test dataset. (d) LFSD dataset.

and the car in the sixth last image), the backgrounds or nonsalient objects are not effectively suppressed by the MMCI [65] and PCFN [64] methods (e.g., the box in the sixth image, the Athena sculpture in the seventh image, the cake in the second last image, and the man in the last image).

5) In contrast, the proposed method achieves superior visual performance with highlighted foreground, complete structure, sharp boundaries, and clean background. For example, only the proposed ASIF-Net can locate the salient object in the second last image, which demonstrates the effectiveness and superiority of our proposed ASIF-Net.

Quantitative comparisons, including P–R curves, *F*-measure, MAE score, and *S*-measure are reported in Fig. 5 and Table I. The P–R curve describes the different combination of precision and recall scores, and the closer the PR curve is to the coordinates (1, 1), the better the performance achieves. As shown in Fig. 5, the proposed ASIF-Net achieves both higher precision and recall scores against other compared methods on the NLPR-test and NJUD-test datasets. On the STEREO and LFSD datasets, our method achieves the second-best performance, only slightly lower than the DMRA method [68]. The quantitative measures listed in Table I also demonstrate the superiority of our method with competitive performance. All the measurements achieve the best performance on the NLPR-test and NJUD-test datasets against other comparison methods. Compared with the second-best method, our method achieves the percentage gain of 12.09% in terms of the MAE score on the NLPR-test dataset, and 10.96% in terms of MAE score on the NJUD-test dataset. On the STEREO and LFSD datasets, our method achieves the comparable performance against the best method (i.e., DMRA [68]) under the limited conditions, including a small number of training data, a low-level backbone model, and no pretraining process. In summary, all the quantitative measures demonstrate the effectiveness of the proposed ASIF-Net to some extent.

D. Ablation Studies

To demonstrate the impact of each component of our proposed network on the performance of salient object detection, we use the same network parameters as the aforementioned settings and carry out experiments on the STEREO dataset, involving seven ablation studies.

TABLE I
QUANTITATIVE COMPARISONS OF DIFFERENT METHODS ON FOUR DATASETS. THE BOLD NUMBERS INDICATE THE BEST PERFORMANCE UNDER EACH CASE, WHILE THE UNDERLINE NUMBERS INDICATE THE SECOND-BEST PERFORMANCE UNDER EACH CASE

Method	Backbone	STEREO Dataset			NLPR-Tset Dataset			NJUD-Test Dataset			LFSD Dataset		
		$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
DSG [11]	–	0.7316	0.1744	0.7048	0.6710	0.1592	0.6902	0.6624	0.2055	0.6511	0.7929	0.1717	0.7342
MILPS [12]	–	0.7372	0.1735	0.7159	0.6934	0.1346	0.7291	0.6700	0.2060	0.6653	0.7964	0.1788	0.7370
SMD [13]	–	0.7620	0.1613	0.7398	0.7269	0.1225	0.7336	0.6871	0.1945	0.6775	0.7826	0.1857	0.7381
RCRR [14]	–	0.7360	0.1755	0.7039	0.6710	0.1591	0.6878	0.6401	0.2167	0.6379	0.7810	0.1848	0.7132
DCL [15]	VGG-16	0.8356	0.0947	0.8040	0.7888	0.0691	0.7827	0.7964	0.1200	0.7777	0.7124	0.1759	0.6887
DSS [16]	VGG-16	0.8650	0.0750	0.8172	0.8328	0.0557	0.8125	0.8021	0.1111	0.7701	0.8142	0.1125	0.7811
R3Net [19]	ResNet	0.8106	0.1067	0.7354	0.8320	0.0493	0.8304	0.8052	0.1050	0.7710	0.8413	0.0962	0.7974
ACSD [56]	–	0.7467	0.1840	0.7082	0.6880	0.1560	0.6904	0.7428	0.1904	0.7022	0.8048	0.1829	0.7237
DCMC [58]	–	0.7996	0.1501	0.7190	0.7057	0.1123	0.7104	0.7665	0.1671	0.6903	0.8499	0.1547	0.7456
MBP [55]	–	0.6627	0.1793	0.5574	0.6085	0.1074	0.6123	0.5932	0.2021	0.5297	0.6757	0.2047	0.5688
DF [63]	–	0.6961	0.1738	0.6279	0.6480	0.1079	0.6710	0.6355	0.1987	0.5930	0.7331	0.2018	0.6340
CTMF [47]	VGG-16	0.8265	0.1023	0.8230	0.8407	0.0560	0.8549	0.8572	0.0847	0.8493	0.8147	0.1202	0.7883
PCFN [64]	VGG-16	0.8838	0.0606	0.8702	0.8635	0.0437	0.8592	0.8875	0.0592	0.8768	0.8290	0.1118	0.7919
MMCI [65]	VGG-16	0.8610	0.0796	0.8504	0.8412	0.0591	0.8524	0.8684	0.0789	0.8588	0.8128	0.1318	0.7793
CPFP [67]	VGG-16	–	–	–	0.8878	0.0359	0.8760	–	–	–	0.8498	0.0879	0.8199
DMRA [68]	VGG-19	**0.8953**	**0.0474**	**0.8778**	0.8870	0.0339	0.8646	0.9003	0.0529	0.8804	**0.8723**	**0.0754**	**0.8391**
TANet [69]	VGG-16	0.8865	0.0591	0.8701	0.8765	0.0410	0.8736	0.8882	0.0605	0.8785	0.8275	0.1108	0.7935
ASIF-Net	VGG-16	0.8939	0.0493	0.8686	**0.9002**	**0.0298**	**0.8844**	**0.9007**	**0.0471**	**0.8887**	0.8584	0.0896	0.8144

TABLE II
QUANTITATIVE EVALUATION OF ABLATION STUDIES ON THE STEREO DATASET. THE BOLD NUMBERS INDICATE THE BEST PERFORMANCE UNDER EACH CASE

	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
ASIF-Net w bi	0.8763	0.0563	0.8570
ASIF-Net w/o adv	0.8800	0.0552	0.8598
ASIF-Net w/o igf	0.8706	0.0608	0.8510
ASIF-Net w/o In-b RAM	0.8310	0.0724	0.8246
ASIF-Net w/o im	0.8744	0.0622	0.8515
ASIF-Net w sa	0.8641	0.0824	0.8541
ASIF-Net w ca	0.8794	0.0696	0.8678
ASIF-Net	**0.8939**	**0.0493**	**0.8686**

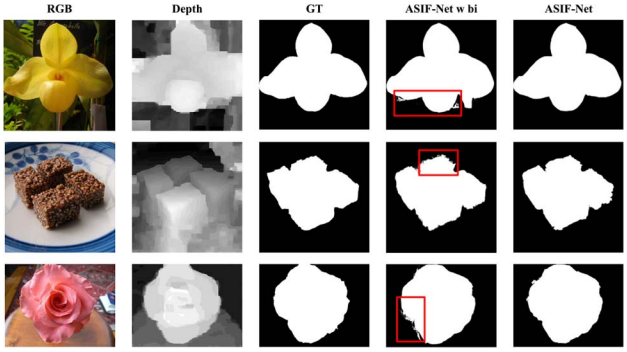

Fig. 6. Visual comparisons between our network using bilinear interpolated side outputs as the attention weights (ASIF-Net w bi) and the proposed network (ASIF-Net). The red boxes indicate the obvious difference.

1) Our network using bilinear interpolated side outputs as the attention weights like [51] (denoted as ASIF-Net w bi).
2) Our network without the adversarial loss (denoted as ASIF-Net w/o adv).
3) Our network without the interaction guidance fusion, which only fuses the RGB-D features in the fusion branch (denoted as ASIF-Net w/o igf).
4) Our network without the In-b RAM (denoted as ASIF-Net w/o In-b RAM).
5) Our network without the identical mapping fashion in the In-b RAM (denoted as ASIF-Net w/o im).
6) The In-b RAM of our network is replaced by the self-attention module [50] (denoted as ASIF-Net w sa).
7) The In-b RAM of our network is replaced by the context-attention module like [75] (denoted as ASIF-Net w ca).

1) ASIF-Net Versus ASIF-Net w bi: In Table II, we can observe that the performance of ASIF-Net w bi is slightly lower than the proposed ASIF-Net, which indicates our side output attention scheme can boost the performance of our method in RGB-D salient object detection. The potential reason is that the side outputs with bilinear interpolation produce the unclear edges and boundaries of the salient objects. Some visual illustrations are shown in Fig. 6. In Fig. 6, compared with the results of the proposed ASIF-Net, the results of ASIF-Net w bi look ambiguous, especially for the edges and boundaries (e.g., the regions in the red boxes). Subjectively, our side output attention scheme is also much more effective than the bilinear interpolation manner used in [51]. It is mainly caused by the operation of bilinear interpolation which uses the distance-weighted average of the nearest pixel values to estimate a new pixel value.

2) ASIF-Net Versus ASIF-Net w/o adv: As reported in Table II, the adversarial learning promotes the performance of saliency detection by constraining the generated salient objects having complete structure and sharp boundary. As shown in Fig. 7, the introduction of adversarial loss can effectively encourage the saliency objects to move toward the complete structure and sharp boundary. For example, the man and the horse are successfully distinguished in the last result of our proposed ASIF-Net.

3) ASIF-Net Versus ASIF-Net w/o igf: Compared with our proposed ASIF-Net, the performance of ASIF-Net w/o igf decreases as reported in Table II, which indicates the interaction guidance is helpful for the fusion of the cross-modal data. The visual validation of the interaction guidance

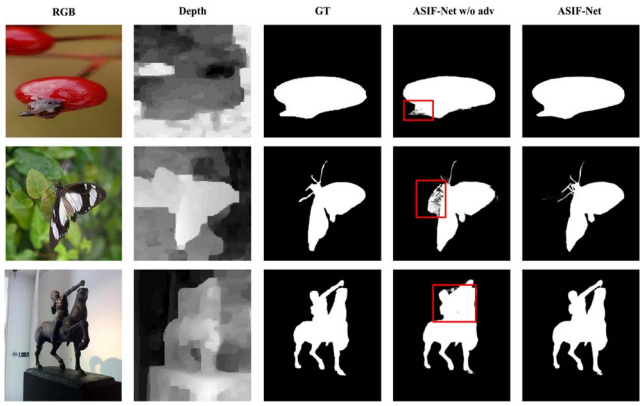

Fig. 7. Visual comparisons between our network without the adversarial loss (ASIF-Net w/o adv) and the proposed network (ASIF-Net). The red boxes indicate the obvious difference.

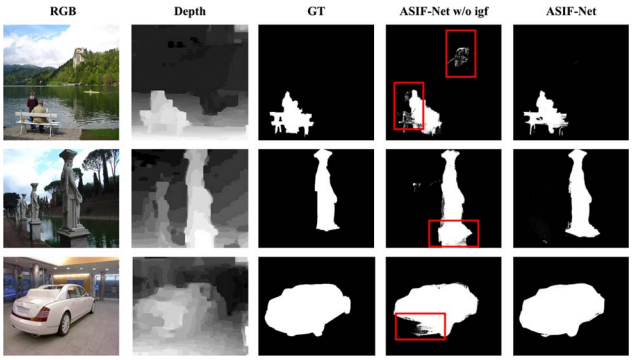

Fig. 8. Visual comparisons between our network without the interaction guidance fusion (ASIF-Net w/o igf) and the proposed network (ASIF-Net). The red boxes indicate the obvious difference.

is shown in Fig. 8. Compared with the clear results of our proposed ASIF-Net, the results of ASIF-Net w/o igf preserve the cluttered background. For example, the mountain in the RGB image in the first row is detected as the salient object in the ASIF-Net w/o igf's result. The main reasons behind this result are that the inconsistency due to the inherent difference existing in the cross-modal data (i.e., RGB image and depth map) and the RGB-D features are not effectively and fully fused.

4) ASIF-Net Versus ASIF-Net w/o In-b RAM: From Table II, it is obvious that the performance of the proposed ASIF-Net significantly decreases when all In-b RAMs are removed (i.e., ASIF-Net w/o In-b RAM). Such a result implies the In-b RAM substantially contributes to the performance of our method. We also provide the visual comparisons in Fig. 9. Without the In-b RAM, the results of ASIF-Net w/o In-b RAM are fragile to the cluttered background regions (e.g., the fence in the first image and the flower bud in the second image), and the shape of salient objects are not clear (e.g., the car in the third image). In contrast, it could be seen that after adding the In-b RAM, our proposed ASIF-Net effectively suppresses these interferences and better highlights the salient objects.

5) ASIF-Net Versus ASIF-Net w/o im: The performance of ASIF-Net without identical mapping (i.e., ASIF-Net w/o im) also drops. As shown in Fig. 10, the saliency maps of ASIF-Net w/o im are incomplete. The main reason may be that the

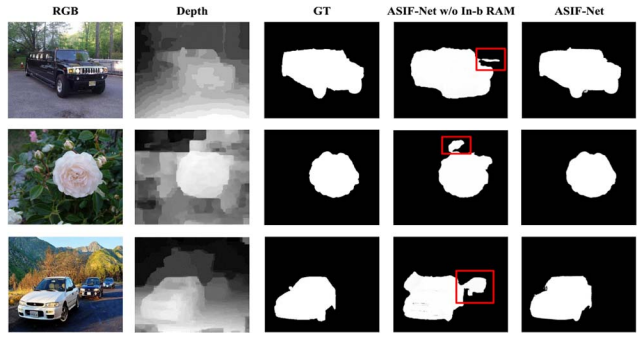

Fig. 9. Visual comparisons between our network without the In-b RAM (ASIF-Net w/o In-b RAM) and the proposed network (ASIF-Net). The red boxes indicate the obvious difference.

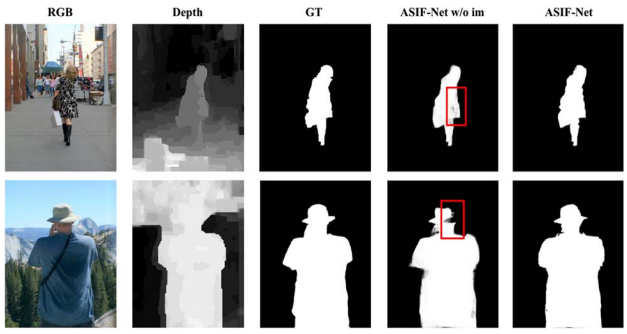

Fig. 10. Visual comparisons between our network without the identity mapping (ASIF-Net w/o im) and the proposed network (ASIF-Net). The red boxes indicate the obvious difference.

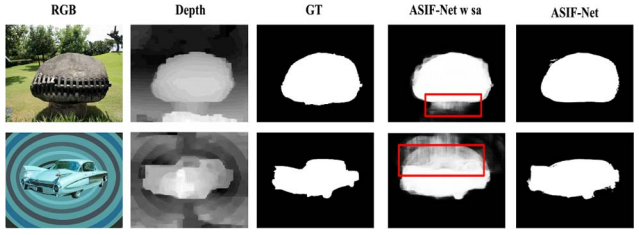

Fig. 11. Visual comparisons between our network with self-attention module (ASIF-Net w sa) and the proposed network (ASIF-Net). The red boxes indicate the obvious difference.

values of the input features become very small or even vanish after multiplying the attention map. In contrast, our ASIF-Net produces more complete the salient objects, and thus achieves better quantitative performance in Table II.

6) ASIF-Net Versus ASIF-Net w sa and ASIF-Net w ca: In terms of the popular attention mechanism, the ASIF-Net with self-attention (i.e., ASIF-Net w sa) and the ASIF-Net with contextual attention (i.e., ASIF-Net w ca) do not achieve better performance in Table II in spite of introducing more parameters. As illustrated in Fig. 11, the results of ASIF-Net w sa bring some backgrounds into the saliency maps due to the relevance between the salient objects and the backgrounds, such as the similar color and depth. This is because of the self-attention mechanism. For the ASIF-Net w ca, its performance is slightly lower than our ASIF-Net in terms of all the measurements. Our ASIF-Net produces sharper edges of salient maps in Fig. 12. The blurring boundaries of ASIF-Net w ca

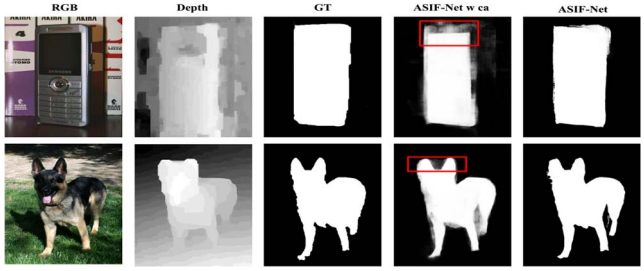

Fig. 12. Visual comparisons between our network with the contextual attention module (ASIF-Net w ca) and the proposed network (ASIF-Net). The red boxes indicate the obvious difference.

is brought by the upsampling on the side outputs of deeper layers following the implementation of [75].

In summary, the proposed ASIF-Net achieves the superior performance to other variant networks in terms of all the evaluation metrics and visual comparisons. Such results prove the effectiveness and contributions of each component of our proposed ASIF-Net.

V. Conclusion

In this article, we presented an ASIF-Net for RGB-D salient object detection. Our method integrates the attention steered complementarity from RGB-D images and introduces the global semantic constraint by using adversarial learning. By introducing these key components into our proposed RGB-D salient object detection network, our method can accurately locate the salient objects and the detected salient objects have complete structure and sharp boundary. Experiments on four datasets demonstrate that the proposed method without bells and whistles outperforms the state-of-the-art salient object detection methods both qualitatively and quantitatively in most of the cases. In addition, the ablation studies also demonstrate the effectiveness of each component of the proposed network. Although our proposed ASIF-Net achieves the superior performance to other variant networks on four benchmarks, it still has some limitations for the challenging RGB-D salient object detection. First, we treat the importance of features from different modalities as the same, which may be inappropriate in some cases, such as the depth map with poor quality induced by the imaging devices or conditions. Second, we do not consider the potential redundant information in the fusion branch. In terms of these two issues, we leave them as future work.

References

[1] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detectioin with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol*, vol. 28, no. 10, pp. 2941–2959, Oct. 2019.

[2] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," 2019. [Online]. Available: arXiv:1904.09146.

[3] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9156–9166, Jun. 2019.

[4] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[5] C.-Y. Li, J.-C. Guo, R.-M. Cong, Y.-W. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5664–5677, Dec. 2016.

[6] X. Cao, C. Zhang, H. Fu, X. Guo, and Q. Tian, "Saliency-aware nonparametric foreground annotation based on weakly labeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1253–1265, Jun. 2016.

[7] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 8, pp. 2014–2027, Aug. 2017.

[8] Q. Jiang, F. Shao, W. Lin, K. Gu, G. Jiang, and H. Sun, "Optimizing multistage discriminative dictionaries for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2035–2048, Aug. 2018.

[9] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. ICCV*, 2013, pp. 2976–2983.

[10] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. CVPR*, 2014, pp. 2814–2821.

[11] L. Zhou, Z. Yang, Z. Zhou, and D. Hu, "Salient region detection using diffusion process on a two-layer sparse graph," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5882–5894, Dec. 2017.

[12] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1911–1922, Apr. 2017.

[13] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[14] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1311–1322, Mar. 2018.

[15] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. CVPR*, 2016, pp. 478–487.

[16] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proc. CVPR*, 2017, pp. 5300–5309.

[17] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. AAAI*, 2018, pp. 6943–6950.

[18] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. CVPR*, 2018, pp. 3089–3098.

[19] Z. Deng *et al.*, "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. IJCAI*, 2018, pp. 684–690.

[20] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jägersand, "BASNet: Boundary-aware salient object detection," in *Proc. CVPR*, 2019, pp. 7479–7489.

[21] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J.-F. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. ICCV*, 2019, pp. 8779–8788.

[22] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

[23] R. Cong *et al.*, "An iterative co-saliency framework for RGBD images," *IEEE Trans. Cybern.*, vol. 49, no. 1, pp. 233–246, Jan. 2019.

[24] Z. Liu, W. Zou, L. Li, L. Shen, and O. L. Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 88–92, Jan. 2014.

[25] Y. Zhang, L. Li, R. Cong, X. Guo, H. Xu, and J. Zhang, "Co-saliency detection via hierarchical consistency measure," in *Proc. ICME*, 2018, pp. 1–6.

[26] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 568–579, Feb. 2018.

[27] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based for co-saliency detection framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, Oct. 2018.

[28] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling, "HSCS: Hierarchical sparsity based co-saliency detection for RGBD images," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1660–1671, Jul. 2019.

[29] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. CVPR*, 2015, pp. 2994–3002.

[30] L. Wei, S. Zhao, O. Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," in *Proc. IJCAI*, 2017, pp. 3041–3047.

[31] D. Zhang, D. Meng, C. Lia, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in *Proc. ICCV*, 2015, pp. 594–602.
[32] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.
[33] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.
[34] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2552–2564, Aug. 2015.
[35] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial–temporal fusion and low-rank coherency diffusion," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3156–3170, Jul. 2017.
[36] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, Dec. 2017.
[37] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4819–4831, May 2019.
[38] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.
[39] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 349–364, Jan. 2018.
[40] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. CVPR*, 2019, pp. 8554–8564.
[41] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. CVPR*, 2014, pp. 2806–2813.
[42] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. CVPR*, 2015, pp. 5216–5223.
[43] T. Wang, Y. Piao, X. Li, L. Zhang, and H. Lu, "Deep learning for light field saliency detection," in *Proc. ICCV*, 2019, pp. 8838–8848.
[44] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2545–2557, Oct. 2019.
[45] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proc. ECCV*, 2018, pp. 135–150.
[46] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. ECCV*, 2014, pp. 92–109.
[47] 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.
[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[49] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[50] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[51] S. Chen, B. Wang, X. Tan, and X. Hu, "Embedding attention and residual network for accurate salient object detection," *IEEE Trans. Cybern.*, early access, doi: 10.1109/TCYB.2018.2879859.
[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012, pp. 1097–1105.
[53] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. CVPR*, 2012, pp. 454–461.
[54] H. Song, Z. Liu, H. Du, G. Sun, O. L. Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, Sep. 2017.
[55] C. Zhu and G. Li, "A multilayer backpropagation saliency detection algorithm and its applications," *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 25181–25197, 2018.
[56] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Process. Image Commun.*, vol. 38, pp. 115–126, Oct. 2015.
[57] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. CVPR*, 2016, pp. 2343–2350.
[58] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819–823, Jun. 2016.
[59] 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.*, early access, doi: 10.1109/TCYB.2019.2932005.
[60] C. Li, C. Guo, J. Guo, P. Han, H. Fu, and R. Cong, "PDR-Net: Perception-inspired single image dehazing network with refinement," *IEEE Trans. Multimedia*, early access, doi: 10.1109/TMM.2019.2933334.
[61] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.
[62] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107038.
[63] 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.
[64] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. CVPR*, 2018, pp. 3051–3060.
[65] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multiscale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 86, pp. 376–385, Jun. 2019.
[66] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.
[67] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proc. CVPR*, 2019, pp. 3927–3936.
[68] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. ICCV*, 2019, pp. 7254–7263.
[69] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Oct. 2015.
[70] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. ICCV*, 2017, pp. 4548–4557.
[71] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversaial networks," 2015. [Online]. Available: arXiv:1511.06434.
[72] S. Karen and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: arXiv:1409.1556.
[73] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. ECCV*, 2014, pp. 345–360.
[74] P. Krahenbuhl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. NeurIPS*, 2011, pp. 109–117.
[75] P. Zhang, L. Wang, D. Wang, H. Lu, and C. Shen, "Agile Amulet: Real-time salient object detection with contextual attention," 2018. [Online]. Available: arXiv:1802.06960.

Chongyi Li received the Ph.D. degree from the School of Electrical and Information Engineering, Tianjin University, Tianjin, China, in 2018.

From 2016 to 2017, he was a joint-training Ph.D. student with Australian National University, Canberra, ACT, Australia. From 2018 to 2020, he was a Postdoctoral Research Fellow with the Department of Computer Science, City University of Hong Kong, Hong Kong. He is currently a Postdoctoral Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research focuses on image processing, computer vision, and deep learning, particularly in the domains of image restoration and enhancement and salient object detection.

Dr. Li is a Guest Editor of special issues for *Signal Processing: Image Communication*.

Runmin Cong (Member, IEEE) received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2019.

He is currently an Associate Professor with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include computer vision and intelligent video analysis, multimedia information processing, saliency detection and segmentation, remote sensing image interpretation, and deep learning.

Dr. Cong was a recipient of the Best Student Paper Runner-Up at IEEE ICME in 2018 and the Excellent Doctoral Dissertation Award from China Society of Image and Graphics. He is a Guest Editor of special issues for *Signal Processing: Image Communication*.

Sam Kwong (Fellow, IEEE) received the B.S. degree in electrical engineering from the State University of New York at Buffalo, Buffalo, NY, USA, in 1983, the M.S. degree from the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree from the University of Hagen, Hagen, Germany, in 1996.

From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada, Ottawa, ON, Canada. He joined Bell Northern Research Canada, Ottawa, ON, Canada, as a member of the Scientific Staff. In 1990, he became a Lecturer with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, where he is currently a Chair Professor with the Department of Computer Science. His research interests are video and image coding and evolutionary algorithms.

Junhui Hou (Member, IEEE) received the B.Eng. degree in information engineering (Talented Students Program) from the South China University of Technology, Guangzhou, China, in 2009, the M.Eng. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2012, and the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2016.

He has been an Assistant Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong, since 2017. His research interests fall into the general areas of visual signal processing, such as image/video/3-D geometry data representation, processing and analysis, semisupervised/unsupervised data modeling for clustering/classification, and data compression and adaptive transmission.

Dr. Hou was a recipient of the Prestigious Award from the Chinese Government for Outstanding Students Study Abroad, China Scholarship Council in 2015, and the Early Career Award from the Hong Kong Research Grants Council in 2018. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *The Visual Computer* and an Area Editor for *Signal Processing: Image Communication*.

Huazhu Fu (Senior Member, IEEE) received the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2013.

From 2013 to 2015, he worked as a Research Fellow with Nanyang Technological University, Singapore, and from 2015 to 2018, he worked as a Research Scientist with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. He is currently a Senior Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include computer vision, image processing, and medical image analysis.

Dr. Fu is an Associate Editor of IEEE ACCESS and *BMC Medical Imaging*.

Guopu Zhu (Senior Member, IEEE) received the B.S. degree in transportation from Jilin University, Changchun, China, in 2002, and the M.S. and Ph.D. degrees in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2004 and 2007, respectively.

He was a Postdoctoral Fellow with Sun Yat-sen University, Guangzhou, China, and a Senior Research Associate with the City University of Hong Kong, Hong Kong. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, where he is the Director of the Center for Internet of Things Computing. He has authored or coauthored more than 30 papers in peer-reviewed international journals. His main research areas are multimedia security, image processing, and control theory.

Prof. Zhu serves as an Associate Editor for *IET Electronics Letters* and the *Journal of Information Security and Applications*. He is a member of the Youth Innovation Promotion Association of Chinese Academy of Sciences.

Dingwen Zhang received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2018.

He is currently an Associate Professor with the School of Machine-Electronical Engineering, Xidian University, Xi'an. From 2015 to 2017, he was a Visiting Scholar with the Robotic Institute, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include computer vision and multimedia processing, especially on saliency detection, video object segmentation, and weakly supervised learning.

Qingming Huang (Fellow, IEEE) received the bachelor's degree in computer science and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is a Chair Professor with the University of Chinese Academy of Sciences, Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He has published more than 400 academic papers in prestigious international journals, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON MULTIMEDIA, and top-level conferences, such as ACM Multimedia, ICCV, CVPR, NIPS, IJCAI, and VLDB. His research areas include multimedia computing, image processing, computer vision, and pattern recognition.

Prof. Huang is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Acta Automatica Sinica*, and a Reviewer of various international journals, including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He has served as the General Chair, the Program Chair, the Track Chair, and a TPC Member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PCM, and PSIVT.