ICNet: Information Conversion Network for RGB-D Based Salient Object Detection

Gongyang Li, Zhi Liu, *Senior Member, IEEE*, and Haibin Ling

***Abstract*— RGB-D based salient object detection (SOD) methods leverage the depth map as a valuable complementary information for better SOD performance. Previous methods mainly resort to exploit the correlation between RGB image and depth map in three fusion domains: input images, extracted features, and output results. However, these fusion strategies cannot fully capture the complex correlation between the RGB image and depth map. Besides, these methods do not fully explore the cross-modal complementarity and the cross-level continuity of information, and treat information from different sources without discrimination. In this paper, to address these problems, we propose a novel Information Conversion Network (ICNet) for RGB-D based SOD by employing the siamese structure with encoder-decoder architecture. To fuse high-level RGB and depth features in an interactive and adaptive way, we propose a novel Information Conversion Module (ICM), which contains concatenation operations and correlation layers. Furthermore, we design a Cross-modal Depth-weighted Combination (CDC) block to discriminate the cross-modal features from different sources and to enhance RGB features with depth features at each level. Extensive experiments on five commonly tested datasets demonstrate the superiority of our ICNet over 15 state-of-the-art RGB-D based SOD methods, and validate the effectiveness of the proposed ICM and CDC block.**

***Index Terms*— RGB-D based salient object detection, information conversion, cross-modal depth-weighted combination, siamese structure.**

I. Introduction

SALIENT object detection (SOD) [1]–[5] aims to automatically identify the most visually attractive object(s) in a scene. It is a fundamental problem on computer vision and plays an important role in many applications such as image/video segmentation [6]–[10] and visual tracking [11], [12]. Recently, depth maps are available due to the large availability of depth sensors such as Microsoft Kinect. Depth information provides complementary geometrical knowledge over RGB information to improve SOD performance, resulting numerous RGB-D based SOD methods [13]–[19].

Manuscript received September 14, 2019; revised January 6, 2020; accepted February 22, 2020. Date of current version March 9, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61771301. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Khan M. Iftekharuddin. *(Corresponding author: Zhi Liu.)*

Gongyang Li and Zhi Liu are with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and also with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: ligongyang@shu.edu.cn; liuzhisjtu@163.com).

Haibin Ling is with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: hling@cs.stonybrook.edu).

Digital Object Identifier 10.1109/TIP.2020.2976689

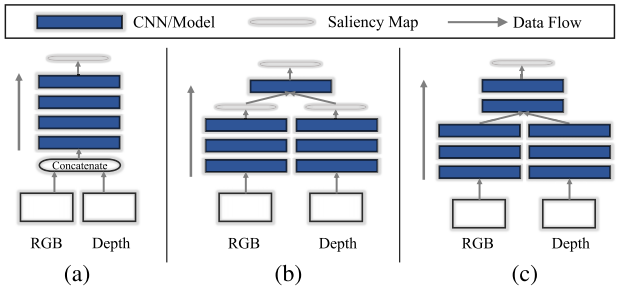


Fig. 1. Three typical fusion architectures to explore the correlation between RGB image and depth map for RGB-D based SOD. (a) Input fusion [20]–[23]; (b) result fusion [13], [24]–[26]; (c) feature fusion [15], [27]–[29].

Obviously, for RGB-D based SOD, it is crucial to effectively fuse RGB image and depth map. Early RGB-D based SOD methods [20]–[23] mainly resort to directly serialize the RGB image and depth map to form a four-channel RGB-D input (Fig. 1(a)) to infer salient object(s). The two-stream based methods have also been proposed for RGB-D based SOD. Specifically, two-stream result fusion based methods [13], [24]–[26] merge two independent saliency maps, which are generated from RGB image and depth map separately, to produce the final saliency map (Fig. 1(b)), while two-stream feature fusion based methods [15], [27]–[29] mainly focus on exploiting complementary features from the separated convolutional neural networks (CNNs) (Fig. 1(c)).

The advantages and shortcomings of these methods are widely known. Input fusion based methods [20]–[23] are the simplest and most straightforward attempts for merging complementary information between RGB image and depth map, but they may not be the most effective. On the contrary, the fusion efficiency of two-stream based methods is slightly better and these methods have achieved appealing advances, but the fusion manner still has room to improve. Result fusion based methods [13], [24]–[26] suffer from the information confusion problem, *i.e.* using summation, multiplication or convolution operation to fuse independent saliency maps without regard to the characteristics of salient objects. Feature fusion based methods [15], [27]–[29] cause the problem of indiscriminate treatment due to treating information from different sources with the same operation during the fusion process. In summary, most current methods aim to improve the effectiveness of fusion, which is important for mining the correlation between RGB image and depth map. However, these methods do not fully explore the cross-modal

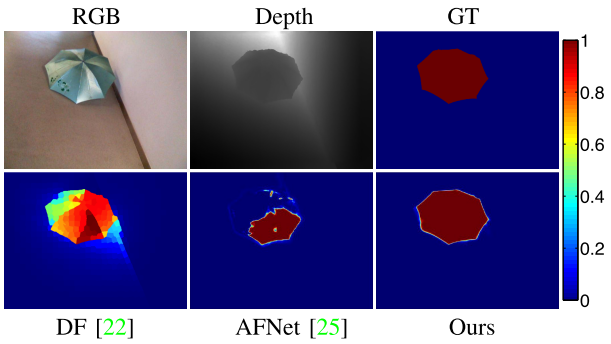


Fig. 2. Visual comparison of two state-of-the-art methods (*i.e.* DF [22] (input fusion) and AFNet [25] (result fusion)) and our proposed method. Comparing with ground truth (GT), our result is clearer and more complete than others.

complementarity and the cross-level continuity of information. There is still a need for a more comprehensive approach to fusing the cross-modal and cross-level information for RGB-D based SOD task.

Through the aforementioned analyses, we propose an effective and efficient network to further promote the efficiency of fusion for RGB-D based SOD in this paper. The proposed network utilizes advantages of the correlation layer [30] and the weighting mechanism to learn common objects location and features enhancement for SOD. Feature fusion (as shown in Fig. 1(c)) is adopted as the basic structure in our network.

To our minds, how to bridge high-level features in the encoder to the decoder in a significative manner is a key problem. High-level features contain sufficient semantic information, which is helpful to recognize and locate the salient objects. Consequently, we propose an *Information Conversion Module* (ICM) to convert high-level features in an interactive and adaptive way. The ICM can recognize the common salient objects among high-level features. Another important issue in designing our network is to make the best use of different levels of cross-modal features in the decoder. In CNNs, features from different sources and different layers make different contributions to SOD, and cross-modal features should not be treated without discrimination. To this end, we propose a *Cross-modal Depth-weighted Combination* (CDC) block to enhance RGB features with depth features, treating features of the two modalities differently. The CDC block is embedded into the decoder to explore cross-modal complementarity and cross-level continuity of features. In this way, the proposed *Information Conversion Network* (ICNet) can better convert the cross-modal and cross-level features for RGB-D based SOD, greatly relieving the shortcomings of previous input fusion, result fusion and feature fusion based methods. Extensive experiments on five challenging datasets demonstrate that the proposed ICNet exhibits competitive performance as compared with 15 state-of-the-art methods in terms of six evaluation metrics. A visual comparison is shown in Fig. 2, comparing with the results of DF [22] (input fusion) and AFNet [25] (result fusion), our saliency map is visually better.

In summary, the contributions of this work are three-fold:

- We propose a novel *Information Conversion Network*, which is equipped with the ICM and the CDC block, for RGB-D based SOD. Our network can automatically learn the optimal conversion of RGB features and depth features and can autonomously determine how to merge them.
- We propose an *Information Conversion Module* for interactively and adaptively exploring the correlation between high-level RGB features and depth features and identifying common salient objects among high-level features.
- We propose a *Cross-modal Depth-weighted Combination Block* to enhance RGB features with depth features at each level, treating RGB features and depth features differently. It not only mines the complementarity of cross-modal features, but also explores the continuity of cross-level features.

The rest of this paper is organized as follows. In Sec. II, we survey the related work. In Sec. III, we present the proposed ICNet for RGB-D based SOD. Extensive experiments are conducted in Sec. IV to compare the proposed ICNet with state-of-the-art RGB-D based SOD methods on five challenging benchmark datasets. The conclusion is given in Sec. V.

II. Related Work

In this section, we briefly introduce several groups of previous work on RGB-D based SOD related to our proposed network.

A. Input Fusion Based Methods for RGB-D Based SOD

Input fusion is an early attempt to mine the correlation between RGB image and depth map in the RGB-D based SOD field, as represented in Fig. 1(a). As its name suggests, input fusion [20]–[23] refers to directly serialize the RGB image and depth map to form a four-channel RGB-D input. Typically, Ren *et al.* [20] adopted the region contrast and surface orientation prior, and refined the result with global optimization. Song *et al.* [21] performed multi-scale pre-segmentation on the RGB-D pair, and proposed the multi-scale discriminative saliency fusion to generate the final saliency map. During the transition from traditional methods to learning-based approaches, a CNN-based input fusion model [22] predicted the salient objects based on the low-level handcrafted saliency feature vectors extracted from the RGB-D pair. Recently, a single stream recurrent CNN [23] was proposed for detecting salient objects with four-channel RGB-D input. In summary, the four-channel RGB-D pair is the most straightforward way to exploit depth map for RGB-D based SOD, but not the most adequate approach.

B. Result Fusion Based Methods for RGB-D Based SOD

Different from the input fusion, some other works called result fusion produce RGB saliency result and depth saliency result independently, and then fuse two results to generate the final saliency map. As shown in Fig. 1(b), result fusion usually adopts the two-stream structure. There were some traditional methods based on hand-crafted features being proposed, especially based on the contrast [13], [24], [31]. Besides, Fang *et al.* [14] extracted the features of color, luminance,

texture, and depth from discrete cosine transform coefficients, and designed an adaptive allocation weight fusion method to combine them. Guo *et al.* [32] iteratively propagated the initial saliency map, which is produced by multiplication, to generate the final saliency map. Considering the quality of the depth map, Cong *et al.* [33] proposed a measure to evaluate the reliability of the depth map, and used it to combine the two predictions. Based on CNNs, Ding *et al.* [26] proposed a saliency fusion network to adaptively fuse the color saliency map and the depth saliency map. Wang *et al.* [25] sent the two saliency maps and a switch map to the saliency fusion module to produce the final result. In general, the summation [13], [14], multiplication [24], [32], [34], [35] and convolution operation [25], [26] are the most commonly used strategies for fusion in result fusion based methods. Although these result fusion based methods have achieved appealing advances, their fusion efficiency is low. The direct summation or multiplication will lead to information confusion, while the convolution operation will lead to inadequate fusion due to not considering the macro-level or micro-level CNN features of salient objects.

C. Feature Fusion Based Methods for RGB-D Based SOD

As shown in Fig. 1(c), the feature fusion, which is also named middle fusion, also adopts the two-stream structure, and it is a very popular solution recently. The biggest advantage of feature fusion is that it not only converts cross-modal features, but also fuses cross-level features, which promotes the performance of RGB-D based SOD by a large margin. Shigematsu *et al.* [27] introduced the deep CNN to RGB-D based SOD for the first time. They inputted ten handcrafted depth features to depth CNN branch, and combined the output features of depth branch with RGB saliency features by concatenation. Finally, they computed saliency scores with two fully connected layers. Han *et al.* [15] proposed an end-to-end transfer network to fuse cross-view features through a new fully connected layer. Although both methods [15], [27] have achieved satisfactory performance, their strategies of feature fusion are straightforward and rough, which makes high-level feature interaction insufficient. The complex feature fusion modes, such as the complementarity-aware fusion module [28], the multi-scale multi-path and cross-modal interaction strategy [29] and the three-stream multi-modal fusion architecture [16], were proposed to further promote complementary information interaction. But whether it is a direct combination operation or a specially designed fusion module, the high-level features are not processed by a proprietary module for effective conversion. On the other hand, since RGB image and depth map are collected from different sources, there is an essential difference between them. But whether in input fusion [23], result fusion [25], [26] or feature fusion [16], [28], [29], these methods handle RGB information and depth information indiscriminately with the same operations, which is unreasonable.

Our ICM can better utilize the high-level features for effective conversion and detecting the common salient objects among them, which establishes a critical bridge between the encoder and the decoder. Moreover, our CDC block treats features of RGB image and depth map differently, which utilizes depth features to assist RGB features, and it can connect features from the encoder to the decoder.

D. Other Methods for RGB-D Based SOD

In addition to these three classic fusion-based methods, there also exist some other methods for RGB-D based SOD. Zhao *et al.* [17] proposed a contrast-enhanced network to produce the one-channel enhanced map in the feature-enhanced module, and designed a fluid pyramid integration method to fuse cross-modal information in a hierarchical manner. Fan *et al.* [19] proposed a depth depurator unit to determine whether the depth information is used in the feature learning module during the prediction process.

III. Proposed Method

In this section, we first describe the overview and motivation of the proposed Information Conversion Network (ICNet) in Sec. III-A. Then we present the Information Conversion Module (ICM) in Sec. III-B. In Sec. III-C, we give the detailed formulas of the Cross-modal Depth-weighted Combination (CDC) block. In the end, we provide implementation details of ICNet in Sec. III-D.

A. Network Overview and Motivation

Our ICNet consists of four components: a siamese encoder for feature extraction, an ICM for information conversion, a CDC block for exploring cross-modal features interaction, and a decoder for cross-level complementary information consolidation and final salient object detection. The overall architecture of our ICNet is illustrated in Fig. 3.

1) Siamese Encoder: Considering the roughness of direct use of depth map and the lack of a suitable pre-trained model, we encode the single-channel depth map to the three-channel HHA [36] representations (*i.e.*, horizontal disparity, height above ground, and the angle of the local surface normal with the inferred gravity direction), which carry more geometrical information than the original single-channel depth map. In order to increase the consistency of two separate encoder networks and to reduce the amount of trainable parameters, we employ the structure of the Siamese network [37] (with the shared weights) as the structure of encoder to extract features from the RGB image and the HHA. Then these pixel-level features are further processed in the ICM and CDC blocks for meaningful information conversion.

Specifically, the siamese encoder is realized by the modified VGG-16 [38], and its parameters are initialized by the well-trained model on ImageNet [39]. There are five-stage features in VGG-16, *i.e.* conv1_2, conv2_2, conv3_3, conv4_3 and conv5_3, which are represented as $\mathbf{F}^R = \{\mathbf{f}_i^R, i = 1, \ldots, 5\}$ of the RGB stream and $\mathbf{F}^D = \{\mathbf{f}_i^D, i = 1, \ldots, 5\}$ of the HHA stream. For efficient computation, the input resolution of the encoder network is set to 288×288, denoted as $W \times H$. Thus, the resolution of features at the ith stage is $[\frac{W}{2^{i-1}}, \frac{H}{2^{i-1}}]$, *i.e.* $[w_i, h_i]$, and the channel number of features at the ith stage is denoted as c_i, *i.e.* $c_i (i = 1, 2, 3, 4, 5) = \{64, 128, 256, 512, 512\}$.

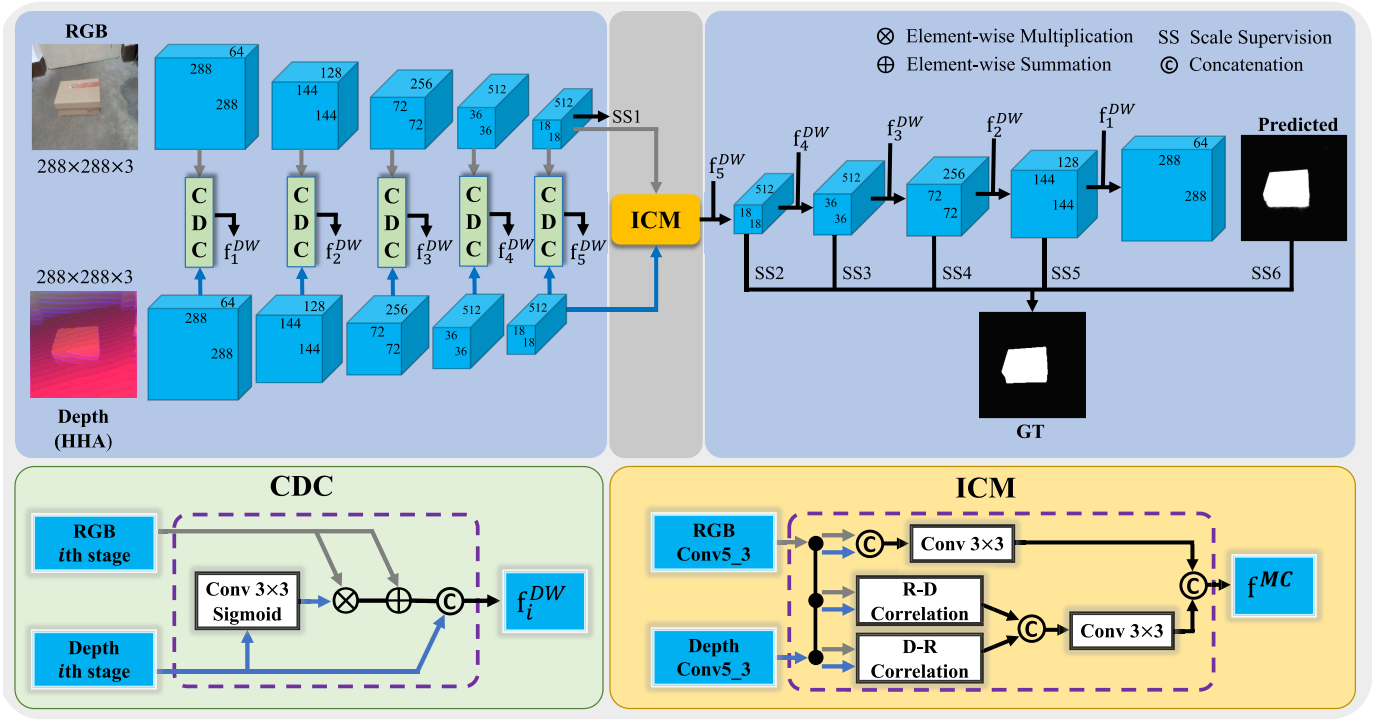


Fig. 3. The overall architecture of the proposed ICNet. We adopt the concatenation-convolution operation and the correlation-convolution operation in the ICM to convert high-level features extracted by the siamese encoder. The output mutual complementarity (MC) features of ICM, *i.e.* $\mathbf{f}^{MC}$, are sent into the decoder network. And five-stage RGB features are also enhanced by the corresponding depth (*i.e.* HHA) features in the CDC block to produce depth weighted (DW) features, *i.e.* $\mathbf{f}_i^{DW}$. Then starting with $\mathbf{f}^{MC}$, $\mathbf{f}_i^{DW}$ gradually participates in the decoder process to infer the salient object(s).

2) ICM Module: The correlation layer [30] is realized by performing multiplicative path comparisons between two feature maps, which is first proposed to discover the displacement of two consecutive video frames for predicting the optical flow. It provides matching capabilities for two feature maps. In the object co-segmentation task [40], the correlation layer is used to detect the common objects between two feature maps, which are from different images. For the RGB-D based SOD, salient objects are not only in the RGB image but also in the depth map. It can be regarded as object co-segmentation between RGB image and depth map to some extent. Inspired by [30], [40], we propose an ICM, which is equipped with the correlation-convolution operation and the concatenation-convolution operation, to precisely highlight the common salient objects among the high-level RGB features and depth features and efficiently convert the sufficient semantic information of high-level CNN features in our ICNet. The detailed explanation of ICM will be described in Sec. III-B. The ICM is a proprietary module for high-level cross-modal features, and it is better than the previous concatenation mode [15], [27], as will be shown in ablation studies in Sec. IV-D.

3) CDC Block: Previous methods [16], [25], [26], [28], [29] treat results/features from different sources (*i.e.* RGB image and depth map) with the same operations/modules, which may be unreasonable. Recently, the attention mechanism is widely used in computer vision field. The spatial attention aims to assign heavy weights to relevant visual parts, but it may be unreasonable to measure many features of different channels with only one attention map. The weighting mechanism can be seen as an upgrade of the spatial attention, and the number of weighting maps is the same as the channel number of features. This makes weighting maps have the ability to assign a weighting map to features of each channel. In our ICNet, to mine the complementary information among features from different sources (*i.e.* RGB image and depth map), we employ features of depth map to enhance features of RGB image via the weighting mechanism, which treats the features of different sources differently. This block will be explained in Sec. III-C in detail. The CDC block provides the depth enhanced RGB features for SOD, and leads to a better SOD performance. The corresponding ablation studies will be shown in Sec. IV-D.

4) Decoder: The ICM has the information conversion ability, and the CDC block has the capability to enhance all the five-stage cross-modal features. To connect these capabilities together, we design a decoder network for salient objects inference process, which explores the continuity of different levels of enhanced features. Inspired by [41], we employ the deep supervision to add pixel-level GT supervision of features in the decoder.

B. Information Conversion Module

High-level CNN features of the siamese encoder network usually contain sufficient semantic and macro-level contextual information, which is helpful to recognize and locate the salient objects. To better convert high-level cross-modal features, we propose an Information Conversion Module (ICM). Concretely, we use the matching capability of the correlation layer to find the common objects between RGB conv5_3 features $\mathbf{f}_5^R$ and depth conv5_3 features $\mathbf{f}_5^D$. Besides, to promote

the robustness of fusion and increase the stability of conversion, we introduce the classical concatenation-convolution operation to merge $\mathbf{f}_5^R$ and $\mathbf{f}_5^D$. In this way, even if the quality of the RGB image or the depth map is unsatisfactory, we can obtain the reasonable converted information by performing the concatenation-convolution operation and the correlation-convolution operation in parallel.

The architecture of ICM is shown in Fig. 3, there are two parallel steps in our ICM. In the first step, the $\mathbf{f}_5^R$ and $\mathbf{f}_5^D$ are concatenated to fed into the convolutional layer with 3×3 kernel size and 512 channels. In the meantime, the $\mathbf{f}_5^R$ and $\mathbf{f}_5^D$ are processed by correlation operations in an interactive and adaptive way, which is $\mathbf{f}_5^R$ to $\mathbf{f}_5^D$ and $\mathbf{f}_5^D$ to $\mathbf{f}_5^R$, *i.e.* "R-D Correlation" and "D-R Correlation" in Fig. 3. Through the mutual correlation operations, both the salient objects on RGB features and the salient objects on depth features can be detected twice, which can avoid missing the salient objects. Then, the interactive correlation features are concatenated to pass through a convolutional layer with 3×3 kernel size and 512 channels for adaptive integration. Finally, features from the concatenation-convolution operation and the correlation-convolution operation are concatenated to generate the mutual complementarity (MC) features, *i.e.* $\mathbf{f}^{MC}$. The process in ICM can be formulated as:

$$\mathbf{f}^{MC} = Conv(\mathbf{f}_5^R + \mathbf{f}_5^D) + Conv(c(\mathbf{f}_5^R, \mathbf{f}_5^D) + c(\mathbf{f}_5^D, \mathbf{f}_5^R)), \quad (1)$$

where $Conv(\cdot)$ represents the convolutional operation with 3×3 kernel size and 512 channels, + denotes the cross-channel concatenation, and $c(\cdot)$ indicates the correlation operation [30] with 31 × 31 patch size and 961 channels.

Specifically, for the ICM, dimensions of both RGB features and depth features, *i.e.* $\mathbf{f}_5^R$ and $\mathbf{f}_5^D$, are (18×18×512)×2, and the dimension of mutual complementarity features, *i.e.* $\mathbf{f}^{MC}$, is 18×18×1024. Both dimensions are the same. And the mutual complementarity features will have stronger complementary expression ability than $\mathbf{f}_5^R$ and $\mathbf{f}_5^D$ without increasing the amount of information.

C. Cross-Modal Depth-Weighted Combination Block

Since RGB image and depth map are collected from different sources, there is an essential difference between them. And in the siamese encoder network, the complementary information exists not only between the high-level CNN features, but also between the low-level CNN features. Inspired by the weighting mechanism, to treat features of RGB image and depth map differently, we propose a Cross-modal Depth-weighted Combination (CDC) block that uses the depth features $\mathbf{f}_i^D$ to assist the RGB features $\mathbf{f}_i^R$. Thus, we can obtain the enhanced features of the salient objects.

As shown in Fig. 3, there are three steps in our CDC block. Firstly, depth features $\mathbf{f}_i^D$ are processed through a convolutional layer with 3×3 kernel size to produce the smooth depth features, which are with the same number of channels as RGB features $\mathbf{f}_i^R$, *i.e.* c_i. These smooth depth features are passed through a Sigmoid function to normalize values to [0,1]. In this way, we get the depth-weight response maps, which can be regarded as the feature-level spatial attention maps and can be used to screen RGB features at feature level. Then, the depth-weight response maps, which are different from the one-channel enhanced map [17] and the spatial attention map, are used to modulate RGB features $\mathbf{f}_i^R$ to focus on the desired part of features by element-wise multiplication. So we get the initial enhanced RGB features. In order to preserve the original RGB information, we also pile the $\mathbf{f}_i^R$ onto the initial enhanced RGB features by a residual connection, *i.e.* element-wise summation, and obtain the final enhanced RGB features. Finally, considering that depth features $\mathbf{f}_i^D$ not only can be used to enhance RGB features $\mathbf{f}_i^R$, but also contains rich information of the salient objects, we combine it with the final enhanced RGB features by cross-channel concatenation and get the depth weighted (DW) features, *i.e.* $\mathbf{f}_i^{DW}$. The process in the CDC block can be represented as:

$$\mathbf{f}_i^{DW} = \mathbf{f}_i^R \oplus (\mathbf{f}_i^R \otimes S(Conv(\mathbf{f}_i^D))) + \mathbf{f}_i^D, \quad (2)$$

where $\oplus$ and $\otimes$ denote the element-wise summation and element-wise multiplication, respectively, and $S(\cdot)$ indicates the Sigmoid function. Thanks to the weighting mechanism in the CDC block, $\mathbf{f}_i^{DW}$ contains rich complementary information with emphasis on RGB information. By using the CDC block at each level, we obtain the depth weighted features at five different levels, which can further promote the SOD.

D. Implementation Details

1) Decoder Network: Since the information continuity exists among the five-level features and the features at different levels have different attributes, *i.e.* the low-level features have edge-aware information and the high-level features have semantic information, we build a decoder network to make full use of these characteristics for gradually inferring the salient objects. As shown in Fig. 3, the structure of the decoder network corresponds to the structure of the encoder network, and we use the deconvolution layer to upsample features to restore resolution in the decoder part. The parameters of the new decoder network are initialized by the xavier method [44]. At each level of the decoder network, $\mathbf{f}_i^{DW}$ is concatenated with the corresponding deconvolution features for successive inference. Notably, we adopt a dropout layer [45] before the deconvolution layer to avoid overfitting.

2) Loss Function: In order to ensure features in the decoder network can accurately reflect the salient objects, we add a convolutional layer with 3×3 kernel size to the back of the RGB encoder stream and each deconvolution layer to obtain the side output saliency maps, and then we adopt the deep supervision [41] with different scales behind the side output saliency maps, *i.e.* "SS1" to "SS6" in Fig. 3. Therefore, the total loss function L_{total} can be represented as:

$$L_{total} = \sum_{k=1}^{6} l_k, \quad (3)$$

where l_k is the softmax loss corresponding to "SS k", and the resolutions corresponding to l_1, l_2, l_3, l_4, l_5 and l_6 are 18×18, 36×36, 72×72, 144×144, 288×288 and 288×288, respectively.

3) Network Training: We implement our network on Caffe [46] in a workstation with an i7-6700K CPU (16GB memory) and a NVIDIA Titan X GPU (12GB memory). Following [17], we randomly select 1400 samples from the NJU2K [42] and 650 samples from the NLPR [24] as the training set. To improve the varieties of training data, we simply augment the training set by mirror reflection and rotation (90°, 180° and 270°), producing 10,250 training triplets totally. We train our network using the standard stochastic gradient descent (SGD) [47] method with dropout ratio 0.5, batch size 1, iteration size 8, momentum 0.9 and weight decay 0.0001. The learning rate is set to 10^{-7} and divided by 10 after 12,500 iterations. We do not use the validation set during the training. The model needs about 25,000 training iterations for convergence, which takes nearly 14.5h.

IV. Experiments

In this section, we first introduce datasets and evaluation metrics in Sec. IV-A and Sec. IV-B, respectively. Then, we compare the proposed information conversion network (ICNet) with state-of-the-art RGB-D based SOD methods in Sec. IV-C. Next, we conduct comprehensive ablation studies to demonstrate the usefulness of the two proposed modules and adopted training strategies in Sec. IV-D. Finally, we present some examples of failure cases and analyze errors in Sec. IV-E.

A. Datasets

In this paper, we have conducted experiments on five widely used public benchmark datasets with different characteristics.

STEREO [31], which is also known as SSB1000, has 1000 pairs of stereo images for testing. These images are mainly collected from the Internet with various resolutions, and the quality of the depth map is relatively coarse.

NJU2K [42] contains 2003 stereoscopic image pairs collected from Internet, daily life and 3-D movies. In order to protect personal privacy, 18 personal images are deleted, leaving 1985 image pairs. It is split into a training set (1400 images), a validation set (100 images), and a testing set (485 images).

LFSD [43] is a relatively small dataset for testing which contains 100 images with depth information captured via a Lytro light field camera and manually labeled ground truths. The resolutions of these images are relatively small.

DES [34] contains 135 RGB-D images from seven indoor scenarios collected by Kinect for testing, which is also named RGBD135. The background of this dataset is relatively simple.

NLPR [24] consists of 1000 images collected by Kinect in 11 different scenes, which includes more than 400 kinds of common objects. It is divided into a training set (650 images), a validation set (50 images), and a testing set (300 images).

B. Evaluation Metrics

We adopt the six most widely used evaluation metrics, *i.e.* classical maximum F-measure (F_β, $\beta^2 = 0.3$), mean absolute error (MAE, $\mathcal{M}$) and precision-recall (PR) curve, and recently proposed weighted F-measure [49], S-measure [50], and maximum E-measure [51], to evaluate the performance of different methods.

1) Weighted F-Measure F_β^w: The weighted F-measure offers a unified solution to the evaluation of non-binary and binary maps. It extends the basic quantities to non-binary values, and weights errors according to their location and their neighborhood. It is formulated as:

$$F_\beta^w = \frac{(1+\beta^2) \times Precision^w \times Recall^w}{\beta^2 \times Precision^w + Recall^w}, \tag{4}$$

where β^2 is set as 1.

2) S-Measure S_α: S-measure is the recently proposed structural similarity measure in the binary map evaluation field. This measure simultaneously evaluates region-aware (S_r) and object-aware (S_o) structural similarity between saliency map and ground truth. And it is defined as:

$$S_\alpha = \alpha * S_o + (1-\alpha) * S_r, \tag{5}$$

where α is the balance parameter and set to 0.5 in this paper.

3) E-Measure E_ξ: The enhanced-alignment measure considers the local pixel-wise values and the image-level mean value together, which is consistent with cognitive vision studies. It is defined as:

$$E_\xi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} f\left(\frac{2\varphi_{GT} \circ \varphi_{FM}}{\varphi_{GT} \circ \varphi_{GT} + \varphi_{FM} \circ \varphi_{FM}}\right), \tag{6}$$

where φ is the bias matrix as the distance between each pixel-wise value of GT and its image-level mean, *i.e.* φ_{GT} and φ_{FM} are for GT and binary foreground map, respectively, and $f(\cdot)$ is a quadratic function.

C. Comparison With the State-of-the-Arts

1) Comparison Methods: We compare our model with 15 state-of-the-art RGB-D based SOD methods including six classical non-deep methods, *i.e.* CDCP [48], ACSD [42], LBE [35], DCMC [33], SE [32] and MDSF [21], and nine CNNs-based methods, *i.e.* DF [22], AFNet [25], CTMF [15], PCF [28], MMCI [29], TANet [16], CPFP [17], DMRA [18] and D3Net [19]. For all the compared methods, we use either the implementations with default parameter settings or the saliency maps provided by the authors for a fair comparison.

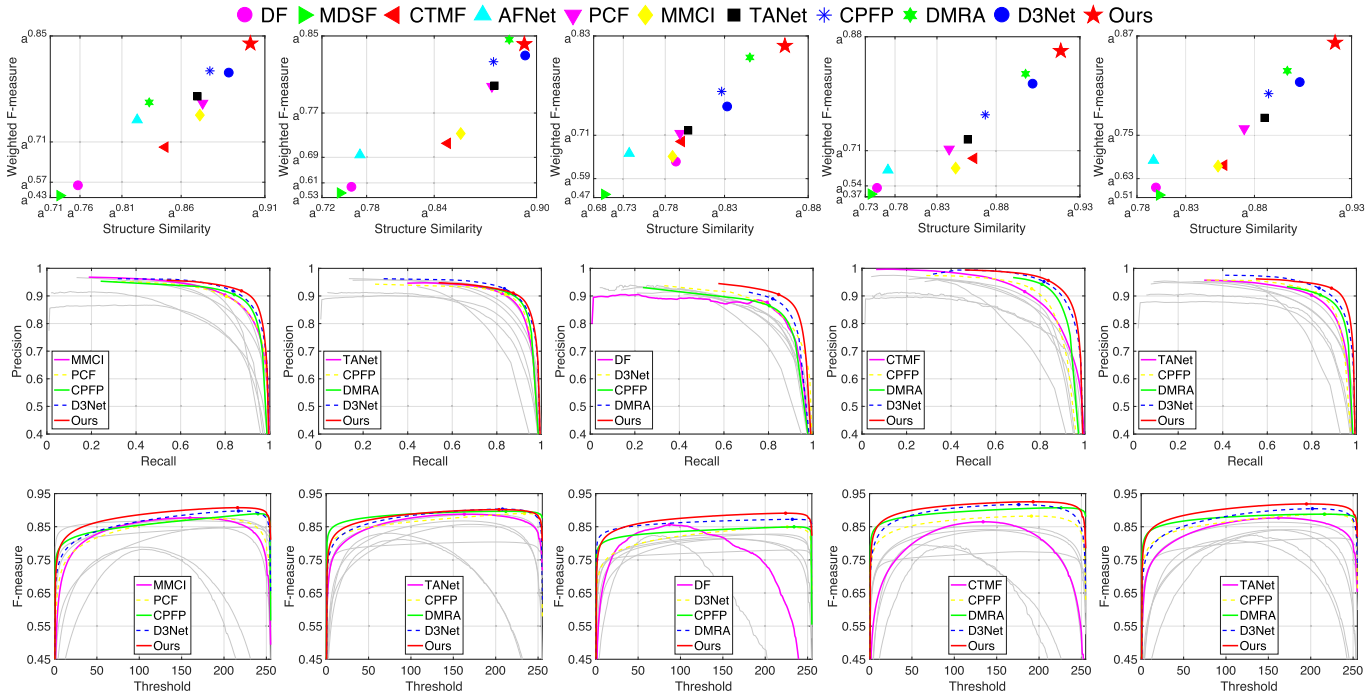
2) Quantitative Performance Comparison: As shown in Table I, our method consistently outperforms all the state-of-the-art methods on four datasets including STEREO [31], LFSD [43], DES [34] and NLPR [24]. On the NJU2K [42] dataset, our method is superior to most methods, except D3Net [19], but our method is also comparable with D3Net. According the *AveRanking* in Table I, our ICNet, D3Net and DMRA are the top 3 methods among five datasets using the four metrics.

We present structure similarity score S_α (X-aixs) and weighted F-measure F_β^w (Y-aixs) of our method and the top 10 state-of-the-art methods [15]–[19], [21], [22], [25], [28], [29] on the 1^{st} row of Fig. 4. Our method shows significant advantages under different evaluation aspects. Besides, the 2^{nd} of Fig. 4 illustrates overall evaluation results of PR curves of our method and top 10 methods on five challenging benchmark datasets. Notably, the lines of top 5 methods are colorful, and the lines of other methods are light gray. Visually, our

TABLE I

QUANTITATIVE SOD RESULTS INCLUDING S-MEASURE, MAXIMUM F-MEASURE, MAXIMUM E-MEASURE AND MAE ON FIVE WIDELY USED DATASETS. THE BEST THREE RESULTS ARE SHOWN IN RED, BLUE, AND GREEN. ↑ & ↓ DENOTE LARGER AND SMALLER IS BETTER, RESPECTIVELY. THE AVERAGE RUNNING TIME (IN SECONDS) OF DIFFERENT METHODS IS ALSO REPORTED. BESIDES, WE PRESENT THE AVERANKING, WHICH IS GENERATED BY CALCULATING THE AVERAGE RANKING OF EACH METHOD USING FOUR METRICS ON FIVE DATASETS

	Metric	CDCP [48]	ACSD [42]	LBE [35]	DCMC [33]	SE [32]	MDSF [21]	DF [22]	AFNet [25]	CTMF [15]	PCF [28]	MMCI [29]	TANet [16]	CPFP [17]	DMRA [18]	D3Net [19]	**ICNet** Ours
	Time	>60.0	0.718	3.110	1.200	1.570	>60.0	10.360	0.030	0.630	0.060	0.050	0.070	0.170	–	0.050	0.075
STEREO [31]	$S_\alpha \uparrow$	0.713	0.692	0.660	0.731	0.708	0.728	0.757	0.825	0.848	0.875	0.873	0.871	0.879	0.835	0.891	0.903
	$F_\beta \uparrow$	0.664	0.669	0.633	0.740	0.755	0.719	0.757	0.823	0.831	0.860	0.863	0.861	0.874	0.847	0.881	0.898
	$E_\xi \uparrow$	0.786	0.806	0.787	0.819	0.846	0.809	0.847	0.887	0.912	0.925	0.927	0.923	0.925	0.911	0.930	0.942
	$\mathcal{M} \downarrow$	0.149	0.200	0.250	0.148	0.143	0.176	0.141	0.075	0.086	0.064	0.068	0.060	0.051	0.066	0.054	0.045
NJU2K-T [42]	$S_\alpha \uparrow$	0.669	0.699	0.695	0.686	0.664	0.748	0.763	0.772	0.849	0.877	0.858	0.878	0.879	0.886	0.895	0.894
	$F_\beta \uparrow$	0.621	0.711	0.748	0.715	0.748	0.775	0.804	0.775	0.845	0.872	0.852	0.874	0.877	0.886	0.889	0.891
	$E_\xi \uparrow$	0.741	0.803	0.803	0.799	0.813	0.838	0.864	0.853	0.913	0.924	0.915	0.925	0.926	0.927	0.932	0.926
	$\mathcal{M} \downarrow$	0.180	0.202	0.153	0.172	0.169	0.157	0.141	0.100	0.085	0.059	0.079	0.060	0.053	0.051	0.051	0.052
LFSD [43]	$S_\alpha \uparrow$	0.717	0.727	0.729	0.753	0.692	0.700	0.791	0.738	0.796	0.794	0.787	0.801	0.828	0.847	0.832	0.868
	$F_\beta \uparrow$	0.703	0.763	0.722	0.817	0.786	0.783	0.817	0.744	0.791	0.779	0.771	0.796	0.826	0.856	0.819	0.871
	$E_\xi \uparrow$	0.786	0.829	0.797	0.856	0.832	0.826	0.865	0.815	0.865	0.827	0.839	0.847	0.872	0.900	0.864	0.903
	$\mathcal{M} \downarrow$	0.167	0.195	0.214	0.155	0.174	0.190	0.138	0.133	0.119	0.112	0.132	0.111	0.088	0.075	0.099	0.071
DES [34]	$S_\alpha \uparrow$	0.709	0.728	0.703	0.707	0.741	0.741	0.752	0.770	0.863	0.842	0.848	0.858	0.872	0.900	0.904	0.920
	$F_\beta \uparrow$	0.631	0.756	0.788	0.666	0.741	0.746	0.766	0.728	0.844	0.804	0.822	0.827	0.846	0.888	0.885	0.913
	$E_\xi \uparrow$	0.811	0.850	0.890	0.773	0.856	0.851	0.870	0.881	0.932	0.893	0.928	0.910	0.923	0.943	0.943	0.960
	$\mathcal{M} \downarrow$	0.115	0.169	0.208	0.111	0.090	0.122	0.093	0.068	0.055	0.049	0.065	0.046	0.038	0.030	0.030	0.027
NLPR-T [24]	$S_\alpha \uparrow$	0.727	0.673	0.762	0.724	0.756	0.805	0.802	0.799	0.860	0.874	0.856	0.886	0.888	0.899	0.906	0.923
	$F_\beta \uparrow$	0.645	0.607	0.745	0.648	0.713	0.793	0.778	0.771	0.825	0.841	0.815	0.863	0.867	0.879	0.885	0.908
	$E_\xi \uparrow$	0.820	0.780	0.855	0.793	0.847	0.885	0.880	0.879	0.929	0.925	0.913	0.941	0.932	0.947	0.946	0.952
	$\mathcal{M} \downarrow$	0.112	0.179	0.081	0.117	0.091	0.095	0.085	0.058	0.056	0.044	0.059	0.041	0.036	0.031	0.034	0.028
AveRanking		14.60	14.15	13.20	12.70	12.30	11.90	9.55	10.10	6.80	6.65	7.25	5.50	3.80	3.20	2.50	1.25



(a) STEREO [31] (b) NJU2K [42] (c) LFSD [43] (d) DES [34] (e) NLPR [24]

Fig. 4. Quantitative comparisons of our method with top 10 state-of-the-art methods on five challenging benchmark datasets. The first row shows the weighted F-measure and the structure similarity score (a = 1000), the second row shows PR curves, and the third row shows F-measure curves.

method is better than other methods, especially on the LFSD and NLPR dataset. F-measure curves are shown in the 3rd row, our method is superior to all other methods, especially on the STEREO, LFSD and NLPR dataset.

Besides, to evaluate whether the improvements of our method are statistically significant or not, we present the t-test analyses between our ICNet and the second best method, *i.e.* D3Net, in Table II. On the four datasets including

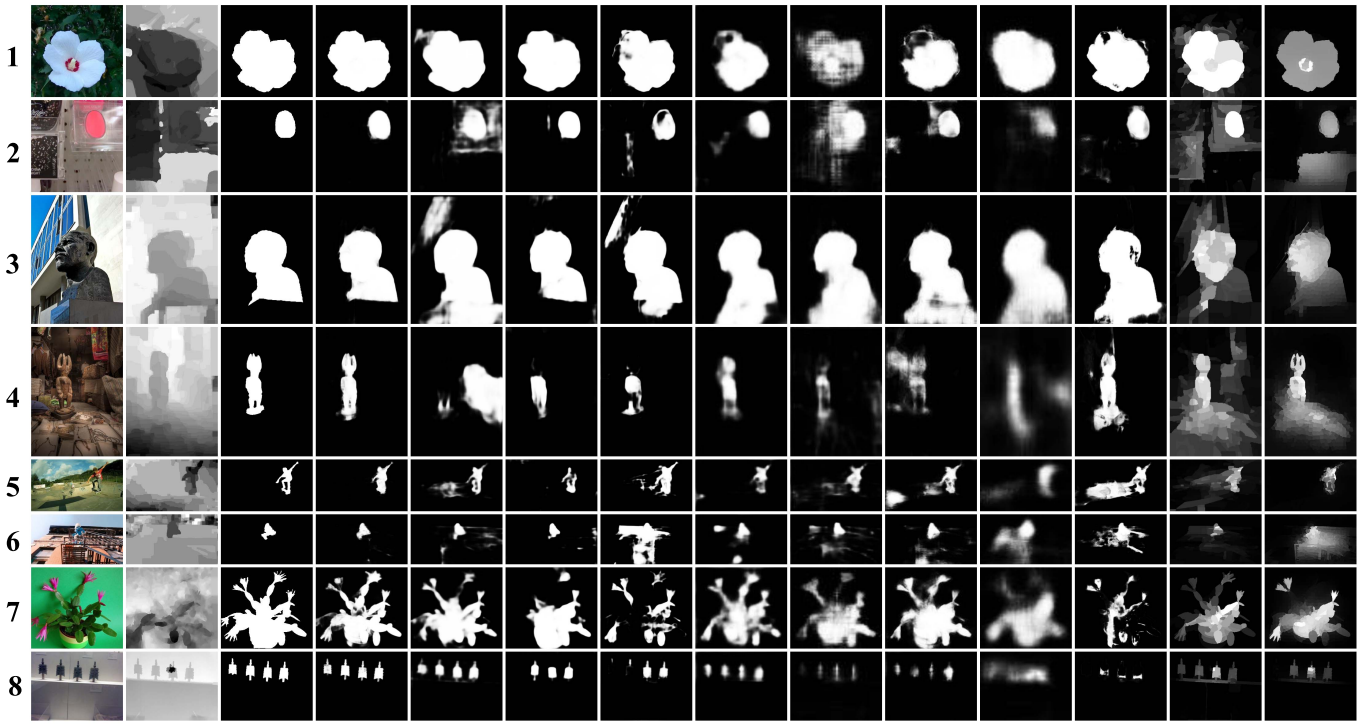


Fig. 5. Visual comparisons to top 10 state-of-the-art methods on different challenging situations.

TABLE II

STATISTICAL ANALYSES OF OUR ICNET AND THE SUBOPTIMAL METHOD D3NET ON FIVE DATASETS. WE PRESENT MEAN AND STANDARD DEVIATION OF EACH METHOD UNDER FOUR METRICS AND CALCULATE THE CORRESPONDING P-VALUES ON T-TEST. THE BEST RESULT IN EACH ROW IS **BOLD**

	Metric	D3Net [19]	**ICNet** (Ours)	P-value
STEREO [31]	$S_\alpha \uparrow$	0.891 ± 0.105	**0.903** ± 0.095	<0.001***
	$F_\beta \uparrow$	0.881 ± 0.166	**0.898** ± 0.138	<0.001***
	$E_\xi \uparrow$	0.930 ± 0.117	**0.942** ± 0.093	<0.001***
	$\mathcal{M} \downarrow$	0.054 ± 0.060	**0.045** ± 0.048	<0.001***
NJU2K-T [42]	$S_\alpha \uparrow$	**0.895** ± 0.109	0.894 ± 0.118	0.940
	$F_\beta \uparrow$	0.889 ± 0.177	**0.891** ± 0.174	0.744
	$E_\xi \uparrow$	**0.932** ± 0.119	0.926 ± 0.124	0.452
	$\mathcal{M} \downarrow$	**0.051** ± 0.057	0.052 ± 0.068	0.855
LFSD [43]	$S_\alpha \uparrow$	0.832 ± 0.155	**0.868** ± 0.127	0.003**
	$F_\beta \uparrow$	0.819 ± 0.225	**0.871** ± 0.172	0.003**
	$E_\xi \uparrow$	0.864 ± 0.179	**0.903** ± 0.142	0.006**
	$\mathcal{M} \downarrow$	0.099 ± 0.099	**0.071** ± 0.073	<0.001***
DES [34]	$S_\alpha \uparrow$	0.904 ± 0.110	**0.920** ± 0.076	0.020*
	$F_\beta \uparrow$	0.885 ± 0.151	**0.913** ± 0.091	0.004**
	$E_\xi \uparrow$	0.943 ± 0.094	**0.960** ± 0.062	0.028*
	$\mathcal{M} \downarrow$	0.030 ± 0.030	**0.027** ± 0.029	0.055
NLPR-T [24]	$S_\alpha \uparrow$	0.906 ± 0.098	**0.923** ± 0.097	<0.001***
	$F_\beta \uparrow$	0.885 ± 0.153	**0.908** ± 0.151	0.003**
	$E_\xi \uparrow$	0.946 ± 0.112	**0.952** ± 0.109	0.257
	$\mathcal{M} \downarrow$	0.034 ± 0.048	**0.028** ± 0.044	0.016*

*p<0.05, **p<0.01, ***p<0.001

STEREO [31], LFSD [43], DES [34] and NLPR [24], we observe that the means of ICNet are bigger than D3Net's, the standard deviations of ICNet are smaller than D3Net's,

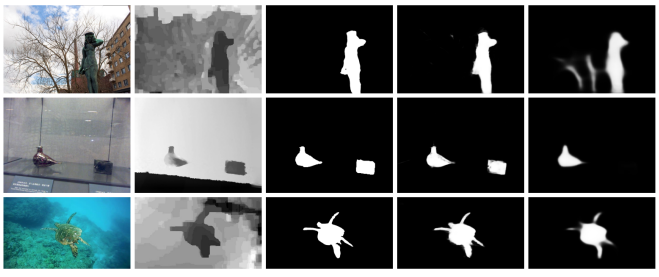


Fig. 6. Visual examples of saliency maps with/without depth map assistance.

and the improvements achieved by ICNet are statistically significant. On the NJU2K [42] dataset, we observe that there is no statistical difference between ICNet and D3Net, which means the performance of ICNet and D3Net is comparable.

3) Qualitative Performance Comparison: In Fig. 5, we show some visualization results of ours and top 10 methods. Specifically, we summarize several challenging situations in RGB-D based SOD: unclear depth (the 2^{nd} row), similar background (the 3^{rd} row), low contrast (the 4^{th} row), complex scene (the 5^{th} row), small object (the 6^{th} row), complex object (the 7^{th} row) and multiple objects (the 8^{th} row).

We show a simple case in the 1^{st} row of Fig. 5 and most of the methods perform well. In the 2^{nd} row, we show some high-contrast images with unclear depth. Due to the fact that the depth information is used to enhance the RGB information in our method, this situation is very challenging for our method. However, our method can still automatically learn the optimal fusion of RGB features and depth features to highlight

TABLE III

ABLATION ANALYSES FOR THE PROPOSED NETWORK ON TWO POPULAR DATASETS STEREO [31] AND NLPR-T [24]. AS CAN BE OBSERVED, EACH COMPONENT IN OUR NETWORK PLAYS AN IMPORTANT ROLE AND CONTRIBUTES TO THE PERFORMANCE. THE BEST RESULT IN EACH COLUMN IS **BOLD**. *w/o Depth*: WITHOUT DEPTH MAP ASSISTANCE, *w/o SW*: WITHOUT SHARED WEIGHTS IN ENCODER, *w/o CON*: WITHOUT CONCATENATION-CONVOLUTION IN ICM, *w/o COR*: WITHOUT CORRELATION-CONVOLUTION IN ICM, *w/o CDC*: WITHOUT CDC BLOCK, *w/o DW*: WITHOUT DEPTH WEIGHTING IN CDC, AND *w/o DS*: WITHOUT DEEP SUPERVISION

Models	STEREO [31]				NLPR-T [24]			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$
ICNet (Ours)	**0.903**	**0.898**	**0.942**	**0.045**	**0.923**	**0.908**	**0.952**	**0.028**
w/o Depth	0.881	0.870	0.927	0.059	0.895	0.876	0.940	0.040
w/o SW	0.898	0.892	0.937	0.048	0.916	0.902	0.948	0.030
w/o COR	0.900	0.894	0.938	0.047	0.914	0.897	0.947	0.031
w/o CON	0.902	0.897	0.941	0.045	0.916	0.900	0.949	0.029
w/o CDC	0.883	0.871	0.931	0.057	0.899	0.874	0.943	0.038
w/o DW	0.900	0.893	0.937	0.047	0.920	0.904	0.952	0.029
w/o DS	0.897	0.892	0.937	0.048	0.916	0.901	0.949	0.030

objects well. The similar background is sometimes associated with objects in the image and is a hard problem in image saliency detection. But in RGB-D saliency detection, the depth information can be utilized to assist the RGB information for saliency detection. In the 3^{rd} row, compared with the previous methods, our method suppresses similar background regions more thoroughly. In the low-contrast images (the 4^{th} row), some methods only segment a part of the salient objects due to the similar colors. In comparison, our method performs very well. In the 5^{th} row, when dealing with images of complex scenes, other methods falsely highlight a lot of background regions, while our method can obtain the clear result. Besides, in the bottom three rows, we also show other three challenging situations related to the object(s), *i.e.* small object, complex object and multiple objects. In these challenging cases, the salient object(s) in our saliency maps are not only complete but also with fine details.

Since MMCI [29] and CTMF [15] roughly reshape the fully connected layer to a low-resolution map, the final saliency maps will be visually blurry. In contrast, our method gradually restores the resolution of features via the encoder-decoder structure, which makes the saliency map very clear and complete.

4) Speed Comparison: We also evaluate the speed performance of different methods, as reported in Table I. For each of the other methods, the average running time is borrowed from [19], which is tested on Titan X GPU with a 224×224 image. The running time of our method is 0.075s, which is tested on Titan X GPU with a 288×288 image. For a fair comparison, we also test the 224×224 image and it only costs 0.063s, which is the upstream level among the running time of these methods. In summary, considering the quantitative performance and the speed, our method is competitive with these state-of-the-art methods.

D. Ablation Studies

In this section, we conduct a more detailed examination of our method on STEREO [31] dataset and NLPR testing set [24]. We evaluate 1) the usefulness of adopting depth map to assist RGB image; 2) the importance of shared weights strategy; 3) the effectiveness of the proposed information conversion module (ICM); 4) the contribution of the proposed CDC block to ICNet; 5) the influence of the extra scale supervision. We change one component each time to assess individual contributions. All the models are retrained with the same hyper-parameters and training set as described in Sec. III-D.

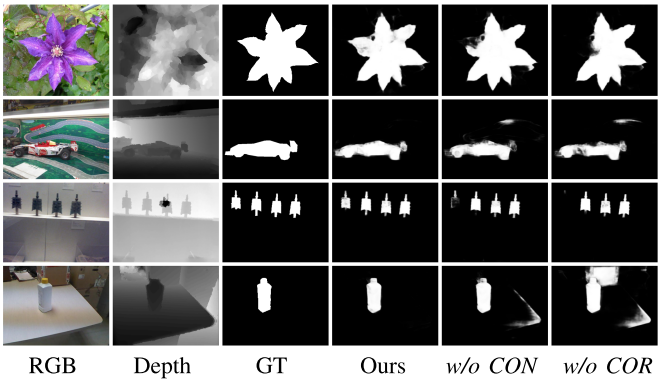


Fig. 7. Saliency maps inferred with different information conversion modes in the ICM. Ours contains both concatenation-convolution operation and correlation-convolution operation. *w/o CON*: without concatenation-convolution, *w/o COR*: without correlation-convolution.

1. Is it useful to adopt depth map to assist RGB image? To validate the usefulness of adopting depth map to assist RGB image in SOD, we delete the depth stream in our ICNet and obtain a basic network without ICM and CDC blocks, named *w/o Depth*. Without the assistance of depth map, we observe that the performance of *w/o Depth* drops sharply, as shown in Table III. To visually demonstrate the usefulness of the depth map, we show several examples of our ICNet and *w/o Depth* in Fig. 6. Since depth map contains geometrical knowledge, we observe that it can assist the RGB image to suppress the similar background, highlight the missing salient object, and provide the details of object.

2. Why using the shared weights strategy? To investigate the importance of using the shared weights strategy in encoder network, we initialize the weights of depth stream by the xavier method and train a new encoder network of depth

stream from scratch, called *w/o SW*. As shown in Table III, we find that the shared weights strategy indeed improves the performance to a certain extent.

3. How effective is the ICM to convert information? To validate that the information conversion module is more effective than the traditional feature fusion mode, *i.e.* the concatenation-convolution operation, we remove the correlation-convolution operation from ICM (namely *w/o COR*), leaving only the traditional feature fusion mode. From Table III, we find that the performance of *w/o COR* is worse than ICNet.

Additionally, to compare the effectiveness of concatenation-convolution operation with correlation-convolution operation, we also only remove the concatenation-convolution operation from ICM (denoted by *w/o CON*). Comparing the results of *w/o COR*, *w/o CON* and ICNet in Table III, we find that the correlation-convolution operation is more effective than the concatenation-convolution operation, and the combination of these two fusion modes, *i.e.* the complete ICM, is better than either of them.

Saliency maps with different information conversion modes are shown in Fig. 7. Comparing the saliency maps in the rightmost three columns with GTs in Fig. 7, we find that ours can better highlight the objects without introducing background regions, and the visualization results of ours are better than any of *w/o CON* and *w/o COR*. Specifically, the depth map in the 1^{st} row of Fig. 7 is confusing, but with the help of ICM, our saliency map is more satisfactory than others. This further demonstrates that the introduction of correlation layer into the ICM is successful and the combination of correlation-convolution and concatenation-convolution is valid for information conversion. As mentioned in Sec. III-B, the input and output of the ICM have the same amount of information, which means that ICM has more powerful information conversion capability without increasing the amount of information.

4. Does the proposed CDC block contribute to ICNet? To evaluate the contribution of the proposed CDC block to ICNet on RGB-D based SOD task, we directly remove the CDC block from the ICNet, and the output features of ICM will be gradually restored to the original resolution via convolution-deconvolution operations without the participation of encoder features from different levels. We report quantitative results in the row with *w/o CDC* of Table III. Comparing with the complete ICNet, the performance of *w/o CDC* drops dramatically. This indicates that the contribution of CDC block is extremely significant and the use of cross-level information continuity is effective.

In addition, in the CDC block, we adopt the weighting mechanism to enhance the RGB features with the depth-weight response maps. To validate the rationality of the depth weighting, we delete the depth weighting in the CDC block, denoted as *w/o DW*, and concatenate $\mathbf{f}_i^R$ and $\mathbf{f}_i^D$ for decoder. The comparison between the results of *w/o CDC* and ICNet in Table III shows that the boost benefits from adding the depth weighting mechanism, and also demonstrates the reasonableness of the design of discriminative treatment in the CDC block.

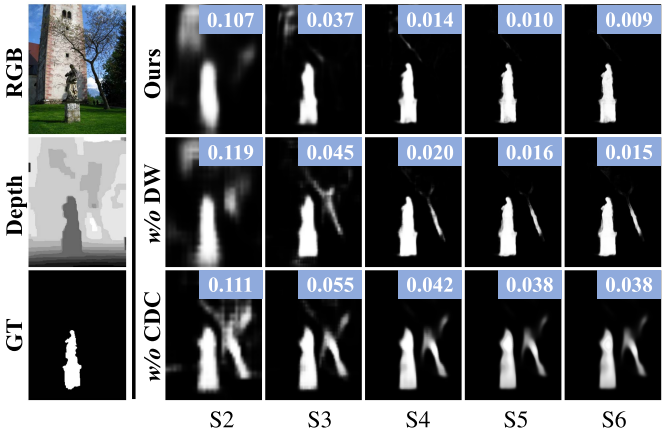


Fig. 8. Side output saliency maps of ours, *w/o DW* and *w/o CDC*. The numbers on the top-right corner of side output saliency maps show the MAE values.

Besides, the side outputs of each layer of ours, *w/o DW* and *w/o CDC* are presented in Fig. 8 to show the capabilities of CDC block intuitively. Visually, without introducing the CDC block (the bottom row in Fig. 8), the side outputs generated from different decoder layers can only basically locate the salient objects without details. To make matters worse, since the cross-modal and cross-level features of the encoder network are not integrated in the decoder, some background regions that reveal in the side outputs of the shallow layers cannot be suppressed in the subsequent decoding process, resulting in a higher MAE.

Even if we directly send the concatenated cross-modal features to the decoder (the middle row in Fig. 8), background regions also cannot be suppressed. Nonetheless, due that the concatenated cross-modal features provide the simple complementary information, *w/o DW* improves the quality of side output saliency maps compared to *w/o CDC*. Furthermore, by embedding the complete CDC block into the decoder, the side output saliency maps of our complete network (the top row in Fig. 8) show a clear effect of refining object and suppressing background. Thanks to the CDC block, which makes a more powerful representation of cross-modal features at different levels and further boosts the decoder, the final output of our method, *i.e.* S6 in the top row, not only accurately highlights salient objects, but also delineates clear object boundaries with a lower MAE.

5. Are the introduced scale loss functions necessary? In the training phase, we adopt the deeply supervised learning mechanism to provide pixel-level supervision and force the intermediate features to have a better representation of salient objects. To study the effect of deep supervision, we remove l_1, l_2, l_3, l_4 and l_5 from Eqn. 3, *i.e.* "SS1" to "SS5" in Fig. 3, and the total loss function L_{total} is only l_6, *i.e.* "SS6" in Fig. 3. Then we retrain our network, and show the results of the variant, namely *w/o DS*, in Table III. We observe that the performance of *w/o DS* is lower than ICNet, and this shows that the introduced scale loss functions are beneficial to our ICNet.

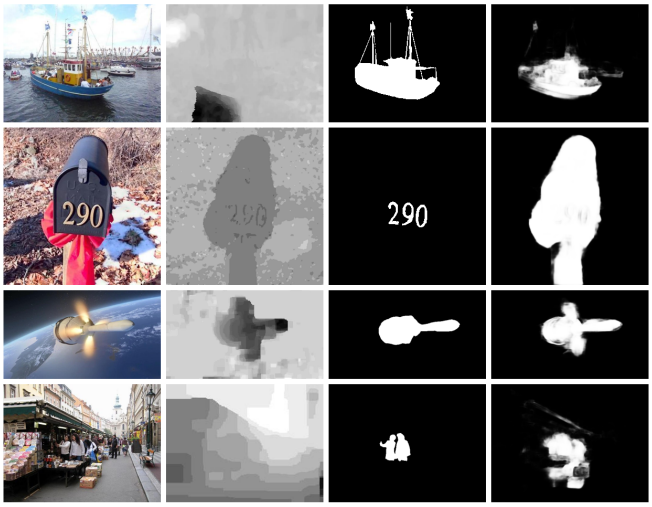


Fig. 9. Some failure cases of our ICNet.

E. Failure Cases and Analyses

As aforementioned, we aim at converting the geometrical prior of depth map to assist RGB image to highlight salient objects in this paper. The quantitative and qualitative evaluations demonstrate the superiority and effectiveness of our ICNet. However, our ICNet fails to produce satisfactory saliency maps when dealing with some scenes such as the examples shown in Fig. 9. In the first example, the boat has a fine structure, such as masts and flags. The corresponding depth map is too fuzzy to provide a prior for the boat. As a result, our ICNet can only roughly highlight the boat without fine details. The second and third examples show the cases with ambiguous annotations. The ground truths of the two examples focus on local object regions, such as the number "290" and the aircraft in Fig. 9. Since our method exploits the prior of depth map, the saliency maps of our ICNet focus on the object regions more globally, *i.e.* the whole mailbox and the aircraft with jet flame in Fig. 9. In the last example, the cluttered scene and the unclear depth map invalidate our ICNet, resulting in an imprecise and blurry result.

Overall, our method relies on the quality of depth map, and an accurate depth map can assist to highlight objects well, as shown in Fig. 6. In other words, a fuzzy depth map can mislead our network to generate an unsatisfactory result, as shown in Fig. 9. Actually, depth maps of existing datasets are collected from various sensors, and the quality of depth maps may be noisy. It is desired to enhance or filter depth maps for further improving the salient object detection performance.

V. Conclusion

In this paper, we propose a novel information conversion network (ICNet). The proposed ICNet consists of two important modules, the information conversion module and the cross-modal depth-weighted combination block. The ICM introduces the concatenation-convolution and correlation-convolution operations to efficiently convert high-level cross-modal features. In the CDC block, the depth features are used to enhance the RGB features, and then this block is embedded into the decoder network for cross-level inference. Experimental results demonstrate that our ICNet can significantly outperform the state-of-the-arts on five challenging benchmark datasets in terms of six evaluation metrics.

For the future work, we think that the proposed information conversion module can be employed for other applications such as video object segmentation and object tracking. On the other hand, RGB features may also be used to assist depth features, which can be embedded in the CDC block to further promote the performance of RGB-D based SOD.

References

[1] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proc. ECCV*, Oct. 2018, pp. 196–212.

[2] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[3] X. Zhou, Z. Liu, C. Gong, and W. Liu, "Improving video saliency detection via localized estimation and spatiotemporal refinement," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2993–3007, Nov. 2018.

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

[5] Y. Liu, D.-P. Fan, G.-Y. Nie, X. Zhang, V. Petrosyan, and M.-M. Cheng, "DNA: Deeply-supervised nonlinear aggregation for salient object detection," 2019, *arXiv:1903.12476*. [Online]. Available: http://arxiv.org/abs/1903.12476

[6] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10494–10503.

[7] G. Li, Z. Liu, and X. Zhou, "Effective online refinement for video object segmentation," *Multimedia Tools Appl.*, vol. 78, no. 23, pp. 33617–33631, Sep. 2019.

[8] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8546–8556.

[9] G. Li, Z. Liu, R. Shi, and W. Wei, "Constrained fixation point based segmentation via deep neural network," *Neurocomputing*, vol. 368, pp. 180–187, Nov. 2019.

[10] Z. Wang, J. Xu, L. Liu, F. Zhu, and L. Shao, "RANet: Ranking attention network for fast video object segmentation," in *Proc. IEEE ICCV*, Oct. 2019, pp. 3978–3987.

[11] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via coarse and fine structural local sparse appearance models," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4555–4564, Oct. 2016.

[12] Z. Chi, H. Li, H. Lu, and M.-H. Yang, "Dual deep network for visual tracking," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2005–2015, Apr. 2017.

[13] X. Fan, Z. Liu, and G. Sun, "Salient region detection for stereoscopic images," in *Proc. 19th Int. Conf. Digit. Signal Process.*, Aug. 2014, pp. 454–458.

[14] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2625–2636, Jun. 2014.

[15] 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.

[16] 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.

[17] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3922–3931.

[18] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE ICCV*, Oct. 2019, pp. 7254–7263.

[19] D.-P. Fan *et al.*, "Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks," 2019, *arXiv:1907.06781*. [Online]. Available: http://arxiv.org/abs/1907.06781

[20] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Y. Yang, "Exploiting global priors for RGB-D saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 25–32.
[21] H. Song, Z. Liu, H. Du, G. Sun, O. Le 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.
[22] 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.
[23] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient object detection for RGB-D image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, Oct. 2019.
[24] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. ECCV*, Sep. 2014, pp. 92–109.
[25] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.
[26] Y. Ding, Z. Liu, M. Huang, R. Shi, and X. Wang, "Depth-aware saliency detection using convolutional neural networks," *J. Vis. Commun. Image Represent.*, vol. 61, pp. 1–9, May 2019.
[27] R. Shigematsu, D. Feng, S. You, and N. Barnes, "Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2749–2757.
[28] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3060.
[29] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 86, pp. 376–385, Feb. 2019.
[30] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
[31] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.
[32] J. Guo, T. Ren, and J. Bei, "Salient object detection for RGB-D image via saliency evolution," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.
[33] 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.
[34] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, Jul. 2014, pp. 23–27.
[35] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2343–2350.
[36] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. ECCV*, Sep. 2014, pp. 345–360.
[37] L. Bertinetto, J. Valmadre, J. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. IEEE ECCVW*, Oct. 2016, pp. 850–865.
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[40] W. Li, O. H. Jafari, and C. Rother, "Deep object co-segmentation," in *Proc. ACCV*, Dec. 2018, pp. 638–653.
[41] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE ICCV*, Dec. 2015, pp. 1395–1403.
[42] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1115–1119.
[43] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE CVPR*, Jun. 2014, pp. 2806–2813.
[44] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, May 2010, pp. 249–256.
[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.
[46] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM MM*, Nov. 2014, pp. 675–678.
[47] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Aug. 2010, pp. 177–186.
[48] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1509–1515.
[49] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.
[50] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.
[51] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

Gongyang Li received the B.E. degree from Shanghai Normal University, Shanghai, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai. His research interests include image/video object segmentation and saliency detection.

Zhi Liu (Senior Member, IEEE) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2005. He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support from the EU FP7 Marie Curie Actions. He has published more than 190 refereed technical articles in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communications. He was a TPC Member/Session Chair of ICIP 2017, PCM 2016, VCIP 2016, ICME 2014, and WIAMIS 2013. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an Area Editor of *Signal Processing: Image Communication*, where he has served as a Guest Editor for the special issue on *Recent Advances in Saliency Models, Applications and Evaluations*.

Haibin Ling received the B.S. and M.S. degrees from Peking University in 1997 and 2000, respectively, and the Ph.D. degree from the University of Maryland, College Park, in 2006. From 2000 to 2001, he was an Assistant Researcher with Microsoft Research Asia. From 2006 to 2007, he worked as a Postdoctoral Scientist at the University of California at Los Angeles. In 2007, he joined Siemens Corporate Research as a Research Scientist. From 2008 to 2019, he worked as a Faculty Member of the Department of Computer Sciences, Temple University. In fall 2019, he joined the Department of Computer Science, Stony Brook University, as a SUNY Empire Innovation Professor. His research interests include computer vision, augmented reality, medical image analysis, and human–computer interaction. He received the Best Student Paper Award at ACM UIST in 2003, the NSF CAREER Award in 2014, the Yahoo Faculty Research and Engagement Program Award in 2019, and the Amazon AWS Machine Learning Research Award in 2019. He has served as the Area Chair for CVPR in 2014, 2016, and 2019. He serves as the Area Chair for CVPR in 2020 and ECCV in 2020. He also serves as an Associate Editors for several journals, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI), *Pattern Recognition* (PR), and *Computer Vision and Image Understanding* (CVIU).