Feature Calibrating and Fusing Network for RGB-D Salient Object Detection

Qiang Zhang, Qi Qin, Yang Yang*, Qiang Jiao*, Jungong Han

***Abstract*—Due to their imaging mechanisms and techniques, some depth images inevitably have low visual qualities or have some inconsistent foregrounds with their corresponding RGB images. Directly using such depth images will deteriorate the performance of RGB-D SOD. In view of this, a novel RGB-D salient object detection model is presented, which follows the principle of calibration-then-fusion to effectively suppress the influence of such two types of depth images on final saliency prediction. Specifically, the proposed model is composed of two stages, i.e., an image generation stage and a saliency reasoning stage. The former generates high-quality and foreground-consistent pseudo depth images via an image generation network. While the latter first calibrates the original depth information with the aid of those newly generated pseudo depth images and then performs cross-modal feature fusion for the final saliency reasoning. Especially, in the first stage, a Two-steps Sample Selection (TSS) strategy is employed to select such reliable depth images from the original RGB-D image pairs as supervision information to optimize the image generation network. Afterwards, in the second stage, a Feature Calibrating and Fusing Network (FCFNet) is proposed to achieve the calibration-then-fusion of cross-modal information for the final saliency prediction, which is achieved by a Depth Feature Calibration (DFC) module, a Shallow-level Feature Injection (SFI) module and a Multi-modal Multi-scale Fusion (MMF) module. Moreover, a loss function, i.e., Region Consistency Aware (RCA) loss, is presented as an auxiliary loss for FCFNet to facilitate the completeness of salient objects together with the reduction of background interference by considering the local regional consistency in the saliency maps. Experiments on six benchmark datasets demonstrate the superiorities of our proposed RGB-D SOD model over some state-of-the-arts.**

***Index Terms*—Salient object detection, RGB-D images, two-steps sample selection, calibration-then-fusion, region consistency aware loss**

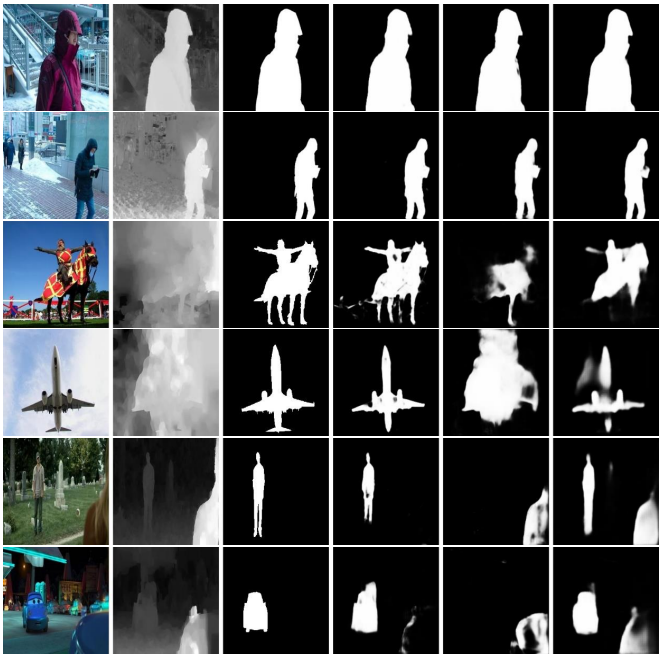
I. Introduction

SALIENT object detection (SOD) imitates the human vision system to identify the most visually appealing objects or regions in an image. It has been widely applied in many computer vision fields, such as object recognition [1], video segmentation [2], person re-identification [3], visual tracking [4] and image quality assessment [5].

Qiang Zhang, Qi Qin, Yang Yang and Qiang Jiao are with Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an, Shaanxi 710071, China, and also with Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China. Email: qzhang@xidian.edu.cn, qinqi67190304@stu.xidian.edu.cn, yang@stu.xidian.edu.cn and qjiao@xidian.edu.cn.

Jungong Han is with Computer Science Department, University of Sheffield, S1 4DP, UK. Email: jungonghan77@gmail.com.

*Corresponding authors: Yang Yang and Qiang Jiao.



(a) (b) (c) (d) (e) (f)

Fig. 1. Visualization of different types of RGB-D images and saliency maps deduced from different input data. (a) RGB images; (b) Depth images; (c) GTs; (d) Saliency maps deduced from RGB images; (e) Saliency maps deduced from depth images; (f) Saliency maps deduced from RGB-D images. In the 1^{st} and 2^{nd} rows, RGB and depth images are both of high visual qualities. In the 3^{rd} and 4^{th} rows RGB images are of high visual qualities, but depth images are of low visual qualities. In the 5^{th} and 6^{th} rows, RGB and depth images contain inconsistent foreground salient objects.

More recently, with the rapid development of Convolutional Neural Networks (CNNs), CNN based SOD models have achieved significant advancements [6–9]. However, these models are mainly designed for visible light images (i.e., RGB images), which are powerless in some challenging or complex scenes [10], [11], e.g., low contrasts between salient objects and backgrounds, cluttered interference, and so on. Different from RGB images that mainly provide some color and texture information, depth images can provide additional geometric structures, such as spatial cues and 3D layouts, which are robust to light and color changing. Compared with RGB information alone, additional depth information has shown the potential to boost the SOD performance. So far, many methods have introduced the depth information into SOD, achieving encouraging results [12–23].

In fact, depth information plays a critical role in RGB-D salient object detection, which directly dictates the performance of subsequent saliency detection. However, depth

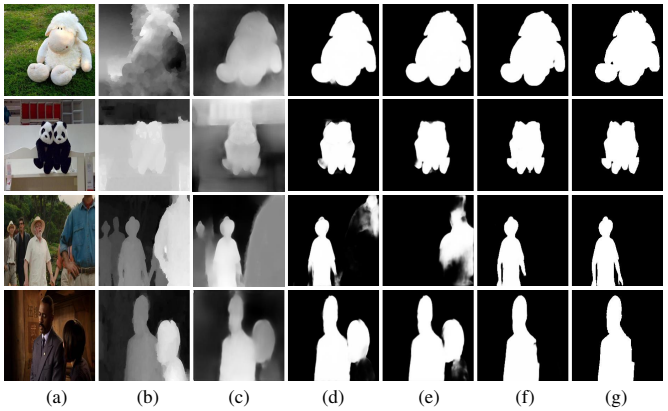


Fig. 2. Visualization of some saliency maps obtained by different models. Here, the depth images are of low visual qualities or have foreground inconsistency with their corresponding RGB images. (a) RGB images; (b) Depth images; (c) Generated pseudo depth images (d) D3Net [24]; (e) CDNet [25]; (f) Ours; (g) GTs.

images are sometimes unreliable. For example, as shown in the 3^{rd} and 4^{th} rows of Fig. 1(b), some depth images have poor visual qualities and contain affluent disturbing cues, which cannot provide much valid spatial information for the cross-modal information fusion. Besides such low-quality depth images, there exists another kind of depth images that also contain unreliable depth cues for RGB-D SOD. As shown in the 1^{st} and 2^{nd} rows of Fig. 1(b) and Fig. 1(e), objects that are closer to the depth camera usually tend to have higher intensities than other objects and are more likely to be regarded as potential salient objects in the depth images. However, as shown in the 5^{th} and 6^{th} rows of Fig. 1(a) and Fig. 1(d), such objects may not be always seen as the salient ones in their corresponding RGB images. Here, we call these depth images foreground-inconsistent ones for brevity in this paper. As shown in Fig.1(f), directly using such low-quality or foreground-inconsistent depth images in RGB-D SOD may contaminate the results of RGB-D SOD.

Recently, some works have paid attention to the qualities of depth images for saliency detection [20], [24–29]. For example, Fan *et al.* [24] designed a depth depurator unit to determine the qualities of depth images and discard those depth images of low visual qualities in the pipeline. Differently, Jin *et al.* [25] first leveraged RGB images to generate some meaningful depth images and then fused such features extracted from the original and those generated depth images for learning robust depth features. Thanks to that, as shown in the 1^{st} and 2^{nd} rows of Fig. 2, these models may work well for those depth images with low visual qualities. However, they may be powerless for those scenes with foreground-inconsistent depth images, as shown in the 3^{rd} and 4^{th} rows of Fig. 2.

As a remedy for such an issue, in this paper, we propose a two-stage RGB-D SOD model to effectively suppress the influence of such two types of depth images on the final saliency prediction via a principle of calibration-and-fusion. Specifically, the proposed RGB-D SOD model is composed of an image generation stage and a saliency reasoning stage. In the image generation stage, we use the input RGB images to generate their corresponding pseudo depth images via an image generation network. In the saliency reasoning stage, we first calibrate such unreliable information in the original depth images with the aid of those generated pseudo depth images and then perform cross-modal feature fusion for the final saliency prediction.

Especially, in the image generation stage, we propose a Two-steps Sample Selection (TSS) strategy to select those high-quality and foreground-consistent depth images from the original RGB-D image pairs as supervision information for the image generation network. More details, in the first step of TSS, such depth images with rich saliency cues will be first selected from the input RGB-D image pairs in terms of the Intersection of Unions (IoUs) between the predicted saliency maps and their corresponding ground truths. On top of that, in the second step, those foreground-inconsistent depth images will be filtered out from such depth images with rich saliency cues according to the true positive rates of their predicted saliency maps. By doing so, such high-quality and foreground-consistent depth images will be selected from the original RGB-D image pairs. This will benefit the generation of some pseudo depth images with more desirable depth information, thus facilitating the subsequent depth information calibration in the saliency reasoning stage.

In the saliency reasoning stage, we propose a novel Feature Calibrating and Fusing network (FCFNet) to effectively calibrate the raw depth information and capture more reliable complementary information from the input RGB-D images for final saliency prediction. More specifically, in FCFNet, a Depth Feature Calibration (DFC) module is first designed to calibrate such unreliable information contained in the original depth images with the aid of the generated pseudo depth images. On top of that, a Multi-modal and Multi-scale Fusion (MMF) module is proposed to capture the cross-modal complementary information and multi-scale context information between the calibrated depth features and the RGB features. As well, in order to make the FCFNet computationally efficient, we just perform MMF on some deeper levels (e.g., the last three levels in this paper) of RGB and calibrated depth features. Accordingly, to avoid the loss of some detailed information contained in the shallower levels of features, which is vital for refining salient object boundaries, a Shallow-level Feature Injection (SFI) module is also presented to inject such detailed information contained in the shallower levels (e.g., the first two levels in this paper) of features into one deeper level (e.g., the third level in this paper) of features in each unimodal feature extraction stream for refining the boundaries of salient objects.

Finally, in addition to the networks, the loss functions are also important for an SOD task. Especially, the Binary Cross Entropy (BCE) loss [30] and the IoU loss [31] are two widely used loss functions in the RGB-D SOD field. However, they are both implemented in a pixel-wise way and ignore the local regional consistency within the saliency maps, thus easily leading to the incomplete detection of salient objects or the introduction of disturbing backgrounds. In view of this, we also design a new loss function, called Region Consistency Aware (RCA), as an auxiliary loss function for our proposed

FCFNet, in which the saliency consistency among the pixels within the foreground salient object regions and the saliency consistency among the pixels within the background regions are simultaneously considered. Under the joint supervision of the three loss functions, i.e., BCE, IoU and RCA, our FCFNet will achieve more accurate saliency results.

Our main contributions are summarized as follows:

(1) We propose a calibration-then-fusion based two-stage model for RGB-D salient object detection, in which the influence of two types of depth images, i.e., low-quality ones and foreground-inconsistent ones, on the saliency detection is simultaneously considered. The results of comprehensive experiments on six benchmark datasets demonstrate the superiorities of our proposed model over existing ones.

(2) In the image generation stage, we propose a TSS strategy to select those high-quality and foreground-consistent depth images from the original input RGB-D image pairs as supervision information for the generation network of pseudo depth images.

(3) In the saliency reasoning stage, we propose a Feature Calibrating and Fusing Network (FCFNet), which is achieved by three dedicated modules, i.e., DFC, MMF and SFI, to first calibrate those unreliable depth information and then capture more cross-modal information from the RGB-D images for final salient object detection.

(4)We design a novel auxiliary loss, i.e., RCA loss, for our proposed FCFNet, in which the local regional consistency within the foreground salient object regions and that within the background regions are simultaneously considered. With the collaboration of BCE, IoU and RCA, more complete foregrounds and less disturbing backgrounds can be achieved by FCFNet.

II. Related Work

A variety of RGB salient object detection (SOD) methods have been proposed in recent years and have achieved outstanding performance [7], [8], [10], [11], [32–36]. However, they rely only on RGB images, which makes them powerless under some challenging scenarios (e.g., transparent regions, cluttered backgrounds and low contrast). For that, depth images, together with RGB images, are introduced to the detection of salient objects, which has been proven to be an effective approach for improving the performance of SOD. In the past few years, various RGB-D SOD models [17–20], [25–27], [37], [38], [39–43] have been proposed to boost the performance of SOD by leveraging both RGB and depth information.

Early RGB-D SOD methods rely on various types of handcrafted features, such as contrast [13] and shape [45], to detect the salient objects. However, these methods usually suffer from unsatisfactory performance due to their limited representation ability of handcrafted features. Recently, CNN-based RGB-D SOD approaches [20], [21], [24], [37], [46] have achieved a qualitative leap in performance due to the powerful feature representation ability of CNNs. Such CNN-based RGB-D SOD approaches may be mainly divided into three categories, i.e., pixel-level fusion based, result-level fusion based and feature-level fusion based ones. Especially, feature-level fusion based RGB-D SOD models have become the current mainstream in the past few years.

Most of these feature-level fusion based RGB-D SOD methods focus on how to effectively integrate cross-modal complementary information from RGB-D images for saliency detection. For example, Chen *et al*. [16] proposed a novel complementarity-aware fusion module to explicitly learn the complementary information from the paired RGB-D images by introducing some cross-modal residual functions and complementarity-aware supervisions. Zhou *et al*. [38] proposed a cross-flow cross-scale adaptive fusion network to detect salient objects in RGB-D images. Chen *et al*. [20] proposed a novel network to explicitly model the potentiality of depth images and effectively integrate the cross-modal complementarity. Cong *et al*. [39] proposed an end-to-end cross-modality interaction and refinement network for RGB-D SOD by fully capturing and utilizing the cross-modality information in an interaction and refinement manner. Xia *et al*. [47] proposed a global contextual exploration network to exploit the role of multi-scale features at a single fine-grained level for RGB-D SOD.

Some recent works have also paid attention to the qualities of input images. For example, Piao *et al*. [27] presented an RGB-D SOD network based on knowledge distillation [12], which can transfer the depth knowledge from the depth stream to the RGB stream, reducing the influence of low-quality depth images on the saliency detection results. Similarly, Chen *et al*. [28] presented a depth quality aware sub-network to evaluate the qualities of depth images before the cross-modal feature fusion. Yang *et al*. [48] presented a Bi-directional Progressive Guidance Network for RGB-D salient object detection, which progressively and interactively suppresses the disturbing cues within the multi-modal input images. Alternatively, Jin *et al*. [25] and Chen *et al*. [26] suggested a promising way to alleviate the influence of low-quality depth images on the detection results by generating some high-quality depth images as complements to the original depth images in RGB-D SOD. Zhang *et al*. [29] exploited some valid priors to alleviate the influence of low-quality depth images for the SOD task. These methods may effectively reduce the influence of low-quality depth images on the final saliency detection. However, they still ignore the influence of such foreground-inconsistent depth images on the RGB-D saliency detection. Differently, in our proposed model, the influence of low-quality depth images and that of foreground-inconsistent depth images on saliency detection are simultaneously considered.

III. Proposed Method

A. Method Overview

The proposed two-stage RGB-D saliency detection model is composed of an image generation stage and a saliency reasoning stage. In the image generation stage, some high-quality and foreground-consistent pseudo depth images will be generated from the input RGB images by using an image generation network. More specifically, a TSS strategy will be employed to select such high-quality and foreground-consistent depth images from the original input RGB-D images

as the supervision information when training the image generation network. In the saliency reasoning stage, a novel FCFNet is designed for SOD, which first calibrates such unreliable depth information in the original depth images with the aid of the generated pseudo depth images and then captures the cross-modal complementary information for the saliency detection. This is achieved by a DFC module, an SFI module and an MMF module. Moreover, together with the BCE and IoU loss functions, a new proposed RCA loss function will be employed to train the FCFNet. In the following contents, we will discuss the two stages in details.

B. Image Generation Stage

As discussed earlier, there may exist some low-quality and foreground-inconsistent depth images in the RGB-D image pairs. Directly using such depth images may degrade the detection performance of RGB-D SOD model greatly. For that, similar to [25], we also perform an image generation network on the input RGB images to generate some high-quality and foreground-consistent pseudo depth images, which will be used to calibrate the original depth images in the subsequent saliency reasoning stage. More details, the same image generation network [49] is employed in this stage to generate pseudo depth images for simplicity, which is composed of several cascaded FCNs and CNNs for discriminative depth estimation.

In addition to the image generation network, the supervision information also plays an important role on the qualities of generated images. Considering that, in this stage, we present a Two-steps Sample Selection (TSS) strategy to select those depth images with high visual qualities and foreground consistency from the input RGB-D image pairs as the supervision information for the image generation network. Fig. 3 illustrates the diagram of the proposed TSS strategy, and the specific details are as follows.

Step1: Select depth samples $\mathbf{D}_a$ containing more saliency cues from the input RGB-D image pairs via the IoUs [31] between the saliency maps predicted by depth images and their ground truths (GTs). This is due to the following considerations. The IoUs can measure the consistency between the depth saliency maps and their corresponding GTs, further reflecting the amount of saliency information contained in the depth images to some extent.

Specifically, we first train an encoder-decoder network to predict the depth saliency maps $\mathbf{P}_d$. Here, the VGG network [50] is employed as the encoder and the decoder part of U-Net [51] is employed as the decoder. After that, for each original depth image (e.g., the k-th depth image), we compute the IoU value $\mathbf{D}_{iou}^k$ between the saliency map $\mathbf{P}_d^k$ and its corresponding ground truth $\mathbf{GT}^k$, i.e.,

$$\mathbf{D}_{iou}^k = \frac{\mathbf{P}_d^k \cap \mathbf{GT}^k}{\mathbf{P}_d^k \cup \mathbf{GT}^k}. \tag{1}$$

Finally, those depth images with more saliency cues, i.e., with $\mathbf{D}_{iou}^k \geq$ th1, are selected from the original RGB-D image pairs, obtaining a new depth image set $\mathbf{D}_a$. Here, the threshold th1 is experimentally set to 0.9 in this paper.

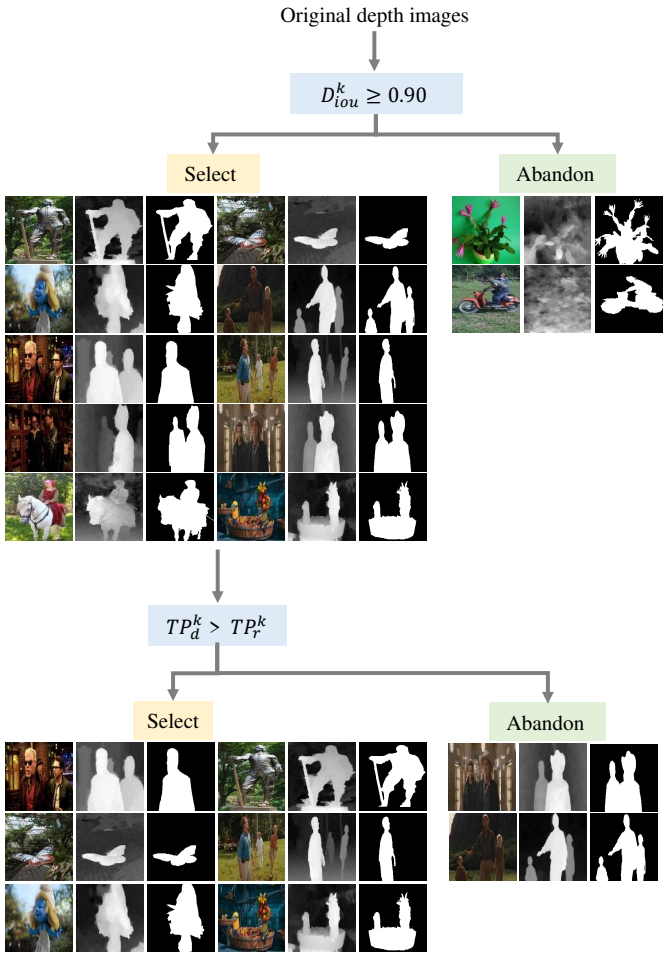


Fig. 3. Diagram of the proposed TSS strategy.

Step2: Select those high-quality and foreground-consistent depth images from $\mathbf{D}_a$ by further computing the true positive rates $\mathbf{TP}_d$ of their predicted saliency maps, obtaining the final depth image set $\mathbf{D}_b$ as supervision information for the image generation network.

Specifically, we first feed RGB images to the re-trained SOD network mentioned in the first step for predicting RGB saliency maps $\mathbf{P}_r$. Then, for each depth image in $\mathbf{D}_a$, we calculate the positive rate $\mathbf{TP}_d^k$ of its depth saliency map $\mathbf{P}_d^k$, as well as the positive rate $\mathbf{TP}_r^k$ of its corresponding RGB saliency map $\mathbf{P}_r^k$, i.e.,

$$\mathbf{TP}_d^k = \frac{\mathbf{P}_d^k \cap \mathbf{GT}^k}{\mathbf{GT}^k}, \mathbf{TP}_r^k = \frac{\mathbf{P}_r^k \cap \mathbf{GT}^k}{\mathbf{GT}^k}. \tag{2}$$

Finally, we select those depth images with $\mathbf{TP}_d^k > \mathbf{TP}_r^k$ from the depth samples set $\mathbf{D}_a$ as the final depth samples set $\mathbf{D}_b$. As shown in Fig. 3, such high-quality and foreground-consistent depth images can be effectively selected from the original RGB-D image pairs by using our proposed TSS strategy. On top of that, we further select those RGB images corresponding to such finally selected depth images from the original RGB-D images, obtaining a new set of RGB-D image pairs $\mathbf{RD}$.

Given the RGB-D image set $\mathbf{RD}$ constructed above, the image generation network in [49] is re-trained in this paper. Here, the selected RGB images are used as the inputs of

(a) (b) (c) (d)

Fig. 4. Visualization of the pseudo depth images obtained by using our proposed TSS strategy. (a) RGB images; (b) Depth images; (c) Generated pseudo depth images; (d) GTs for saliency maps.

the image generation network and their corresponding depth images are used as the supervision information. After training, we feed all of the RGB images contained in the original RGB-D image pairs into the re-trained image generation network to produce their corresponding high-quality and foreground-consistent pseudo depth images. These generated pseudo depth images will be used to calibrate the original depth images in the subsequent saliency reasoning stage. As shown in Fig. 4, compared to those original low-quality or foreground-inconsistent depth images, the generated pseudo depth images usually have higher visual qualities or have better foreground consistency with their corresponding RGB images.

C. Saliency Reasoning Stage

In the saliency reasoning stage, the original depth features are first calibrated by using those pseudo depth images generated from the first stage and then fused with the RGB features for better capturing the cross-modal complementary information within the multi-modal RGB-D images. To this end, we propose a Feature Calibrating and Fusing Network (FCFNet) in this stage for saliency detection

Specifically, as shown in Fig. 5, the proposed FCFNet contains three key components: a DFC module, an MMF module and an SFI module. First, a VGG16 based three-stream encoding network is deployed to simultaneously extract hierarchical features from RGB images, original depth images and pseudo depth images, which are denoted as $\mathbf{r}^{\mathrm{i}}$, $\mathbf{d}_o^i$ and $\mathbf{d}_{pse}^i$ (i = 1,2,3,4,5), respectively. Here, i denotes the feature level index. After that, the DFC module is employed to calibrate such original depth features $\mathbf{d}_o^i$ with the aid of the pseudo depth features $\mathbf{d}_{pse}^i$, obtaining the calibrated depth features $\mathbf{d}^i$. Subsequently, several MMF modules are performed on the RGB features $\mathbf{r}^i$ and the calibrated depth features $\mathbf{d}^i$, achieving cross-modal feature fusion. Especially, considering the trade-off between the computational complexity and the saliency detection performance, the MMF module is only performed on the last three levels as in [52].Meanwhile, to avoid the loss of some detailed information from the shallower levels during the cross-modal feature fusion, the proposed SFI module is further employed to inject such valuable unimodal features from the shallower levels (i.e, the first two levels) into the middle level (i.e., the third level) of unimodal features before they are fused. Finally, the fused cross-modal features are integrated in a progressive way, and some auxiliary loss functions are applied to facilitate the optimization, obtaining three levels of saliency maps $\mathbf{S}^{(t)}$ (t = 3,4,5). Here, $\mathbf{S}^{(3)}$ is taken as the final saliency map in this paper. Especially, the existing BCE, IoU and our proposed RCA loss functions are jointly performed on FCFNet for better saliency detection results. Details about these components mentioned above are seen in the following contents.

*1) **Depth Feature Calibration Module**:* As discussed earlier, there may exist some low-quality or foreground-inconsistent depth images in the original RGB-D image pairs. Directly using the features extracted from the original depth images may easily deteriorate the performance of an RGB-D SOD model. In order to address such a problem, as shown in Fig. 6, we propose a Depth Feature Calibration (DFC) module to calibrate such original depth features with the aid of the generated pseudo depth images before they are fed into the cross-modal feature fusion module.

Specifically, we first initially suppress those unreliable cues contained in original depth features $\mathbf{d}_o^i$ by utilizing the extracted pseudo depth features $\mathbf{d}_{pse}^i$ to re-weight the importance of unimodal features in the channel dimension. More specific, the original depth features $\mathbf{d}_o^i$ and the pseudo depth features $\mathbf{d}_{pse}^i$ are first concatenated to learn a set of channel-wise weights via convolution and global average pooling operations, i.e.,

$$\mathbf{W}^i = \sigma(\mathrm{GAP}(\mathrm{Conv}_3(\mathrm{Cat}(\mathbf{d}_o^i, \mathbf{d}_{pse}^i), \alpha))), \tag{3}$$

where GAP($*$) denotes the global average pooling. $\sigma(*)$ refers to the Sigmoid operation. $\mathbf{W}^i$ denotes the channel-wise weights for the i-th level of original depth features $\mathbf{d}_o^i$. $\mathrm{Conv}_3(*, \alpha)$ denotes a 3×3 convolutional layer with its parameters α. $\mathrm{Cat}(*,*)$ denotes the concatenation operation. With the channel-wise weights, the initially calibrated depth features $\mathbf{d}_{ic}^i$ are obtained by

$$\mathbf{d}_{ic}^i = \mathbf{W}^i \otimes \mathbf{d}_o^i \oplus (\mathbf{1} - \mathbf{W}^i) \otimes \mathbf{d}_{pse}^i, \tag{4}$$

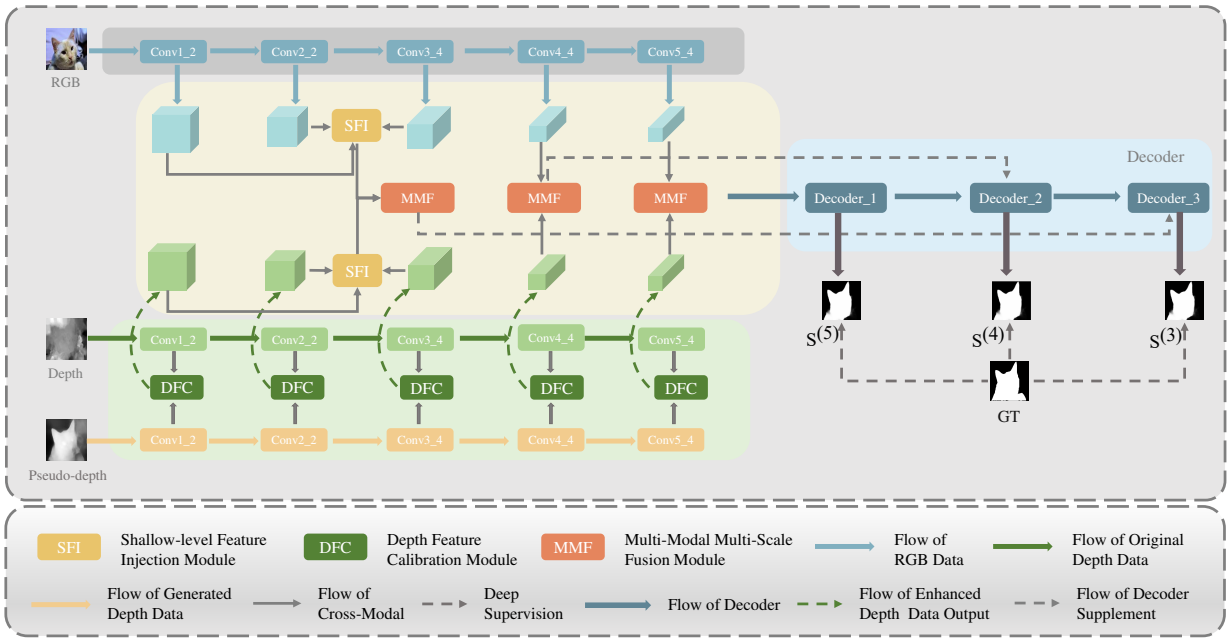


Fig. 5. Diagram of the proposed FCFNet. First, RGB features, original depth features and pseudo depth features are extracted from the three-stream backbone network, respectively. Then, original depth features and pseudo depth features are fed into the DCF module to calibrate those original depth features. Next, unimodal RGB features and calibrated depth features are fused via the proposed MMF module to capture cross-modal complementary information and multi-scale context information. In addition, in order to avoid the loss of some detailed information contained in the shallower levels of features, the SFI module is applied to inject such important detailed information from the first two levels into the third level before performing cross-modal feature fusion. Finally, those cross-modal features are fed into the decoder to progressively achieve saliency reasoning.

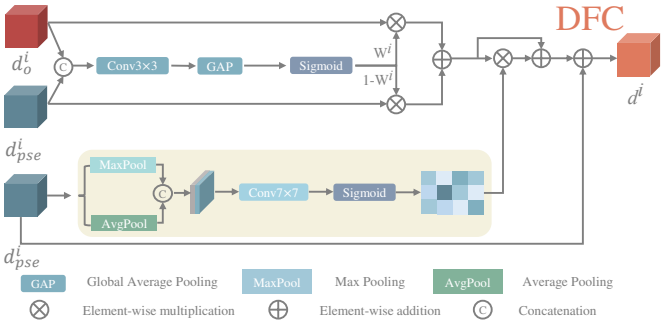


Fig. 6. Architecture of the proposed DFC.

where $\otimes$ denotes the element-wise multiplication and $\oplus$ denotes the element-wise addition. $\mathbf{1}$ denotes an all-one vector with the same size as $\mathbf{W}^i$.

Then, the pseudo depth features $\mathbf{d}^i_{pse}$ are further applied to calibrate $\mathbf{d}^i_{ic}$ via a spatial-attention mechanism, achieving the salient content alignment, i.e.,

$$\mathbf{d}^i_{en} = \mathbf{d}^i_{ic} \otimes \mathbf{W}^i_{\mathrm{sa}}(\mathbf{d}^i_{pse}) \oplus \mathbf{d}^i_{ic}, \tag{5}$$

where $\mathbf{W}_{\mathrm{sa}}(*)$ denotes the spatial weight generation function [53], i.e.,

$$\mathbf{W}^i_{\mathrm{sa}}(\mathbf{d}^i_{pse}) = \sigma(\mathrm{Conv}_7(\mathrm{Cat}(\mathrm{Avg}(\mathbf{d}^i_{pse}), \mathrm{Max}(\mathbf{d}^i_{pse})), \omega)). \tag{6}$$

Here, $\mathrm{Conv}_7(*, \omega)$ denotes a 7×7 convolutional layer with its parameters ω. $\mathrm{Avg}(*)$ denotes the average pooling operation along the channel dimension. $\mathrm{Max}(*)$ denotes the max pooling operation along the channel dimension.

Finally, the pseudo depth features $\mathbf{d}^i_{pse}$ are further embedded into the enhanced depth features $\mathbf{d}^i_{en}$ via a skip connection, obtaining the finally calibrated depth features $\mathbf{d}^i$, i.e.,

$$\mathbf{d}^i = \mathbf{d}^i_{pse} \oplus \mathbf{d}^i_{en}\,. \tag{7}$$

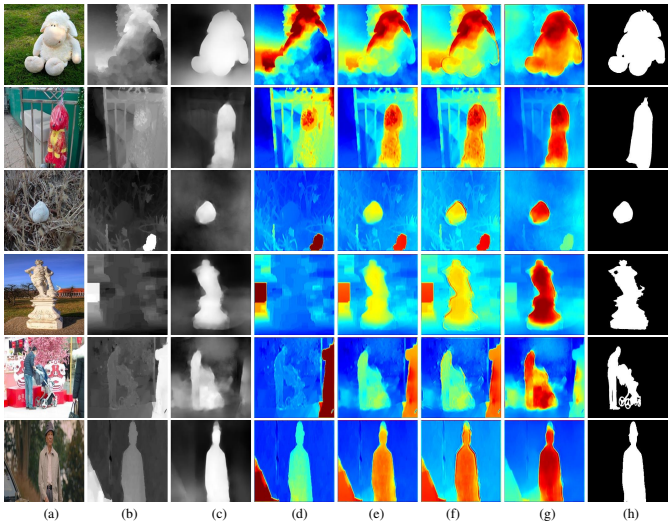


Fig. 7. Visualization of calibrated depth features. (a) RGB images; (b) Original depth images; (c) Generated pseudo depth images; (d) Original depth features $\mathbf{d}^1_o$; (e) Initially calibrated depth features $\mathbf{d}^1_{ic}$; (f) Enhanced depth features $\mathbf{d}^1_{en}$; (g) Finally calibrated depth features $\mathbf{d}^1$; (h) GTs for saliency maps.

Fig. 7 visualizes the calibration of original depth features with the aid of the extracted pseudo depth features in the DFC module. Compared with the original depth features in Fig. 7 (d), the finally calibrated depth features in Fig. 7 (g) contain more reliable depth information about the salient objects, while suppressing disturbing cues within the background regions.

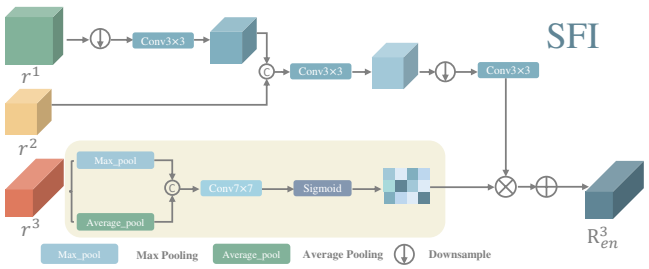


Fig. 8. Architecture of the proposed SFI.

*2) **Shallow-level Feature Injection Module**:* As aforementioned, we only perform the cross-modal feature fusion on the last three levels of unimodal features for reducing computational complexity. However, doing that will lose some important detailed information from shallower levels, which is crucial for refining the boundaries of salient objects. In view of this, we design an SFI module that selectively injects those valuable unimodal features from the first two levels into the third level of unimodal features.

Fig. 8 illustrates the details of the proposed SFI module, where the injection of shallower-level RGB features is taken as an example. As shown in Fig. 8, in SFI, the first two levels of unimodal RGB features are first concatenated to obtain the shallower levels of fusion features $\mathbf{r}_{sl}$, i.e.,

$$\mathbf{r}_{sl} = \mathrm{Conv}_3(\mathrm{Cat}(\mathrm{Conv}_3(\mathrm{Down}(\mathbf{r}^1), \varepsilon), \mathbf{r}^2), \beta), \tag{8}$$

where Down$(*)$ denotes the downsample operation. $\mathrm{Conv}_3(*, \varepsilon)$ and $\mathrm{Conv}_3(*, \beta)$ denote two 3×3 convolutional layers with their parameters ε and β, respectively.

The same spatial-attention mechanism mentioned in the DFC module is then applied to the third level of RGB features $\mathbf{r}^3$ for selecting those valuable detailed information from $\mathbf{r}_{sl}$. Finally, such selected RGB features are further injected into the third level of RGB features, obtaining the enhanced RGB features $\mathbf{r}^3_{en}$. The process can be described as:

$$\mathbf{r}^3_{en} = \mathrm{Conv}_3(\mathrm{Down}(\mathbf{r}_{sl}), \gamma) \otimes \mathbf{W}_{\mathrm{sa}}(\mathbf{r}^3) \oplus \mathbf{r}^3, \tag{9}$$

where $\mathbf{r}^3_{en}$ denotes the third level of enhanced RGB features. $\mathrm{Conv}_3(*, \gamma)$ denotes a 3×3 convolutional layer with its parameters γ. Accordingly, we also obtain the middle level of enhanced depth features $\mathbf{d}^3_{en}$.

*3) **Multi-modal Multi-scale Fusion Module**:* Employing those simple fusion strategies is hard to fully exploit complementary information within RGB-D image pairs. Considering that, we propose a Multi-modal Multi-scale Fusion (MMF) module to achieve the fusion of RGB and calibrated depth features, where the unimodal features ($\mathbf{r}^i$ and $\mathbf{d}^i$) are first fused and then enhanced from a cross-scale perspective for achieving better saliency detection results. As shown in Fig. 9, in MMF, a weighted channel attention mechanism is first employed to re-weight and fuse the corresponding channels of unimodal features from different modalities, resulting in initially fused features. Afterwards, a Multi-Scale Attention (MA) module is designed, where the initially fused features are fed into four parallel branches to capture their multi-scale contextual information for dealing with the challenge of object scale variations in the scenes. Meanwhile, the importance of those extracted multi-scale features is re-assigned to obtain the finally fused features.

Specifically, unimodal features $\mathbf{r}^i$ and $\mathbf{d}^i$ are first concatenated and then fed into some convolution layers to learn the channel-wise relative importance weights, which can be formulated as:

$$\mathbf{F}^i = \mathrm{CB}(\mathrm{Cat}(\mathbf{r}^i, \mathbf{d}^i), \varphi), \tag{10}$$

$$\mathbf{W}^i_{\mathrm{fus}} = \sigma(\mathrm{MLP}(\mathrm{GAP}(\mathbf{F}^i), \eta) + \mathrm{MLP}(\mathrm{GMP}(\mathbf{F}^i), \psi)), \tag{11}$$

where $i = 3, 4, 5$. Especially, when $i = 3$, $\mathbf{r}^3$ and $\mathbf{d}^3$ are the enhanced unimodal features ($\mathbf{r}^3_{en}$ and $\mathbf{d}^3_{en}$) obtained by the SFI module, respectively. $\mathbf{F}^i$ denotes the concatenated features of $\mathbf{r}^i$ and $\mathbf{d}^i$. $\mathrm{CB}(*, \varphi)$ denotes a convolutional blocks with its parameters φ, which contains two 3×3 convolutional layers. $\mathrm{MLP}(*, \eta)$ and $\mathrm{MLP}(*, \psi)$ denote two fully connected blocks with their parameters η and ψ, respectively. $\mathrm{GAP}(*)$ and $\mathrm{GMP}(*)$ denote the global average pooling operation and the global max pooling operation, respectively. $\mathbf{W}^i_{fus}$ denotes a set of channel-wise weights. With the obtained channel-wise weights $\mathbf{W}^i_{fus}$, the fused features are obtained by

$$\mathbf{F}^i_{fus} = \mathbf{r}^i \otimes \mathbf{W}^i_{fus} + \mathbf{d}^i \otimes (\mathbf{1} - \mathbf{W}^i_{fus}), \tag{12}$$

where $\mathbf{F}^i_{fus}$ denotes the i-th level of initially fused features. $\mathbf{1}$ denotes an all-one vector with the same size of $\mathbf{W}^i_{fus}$.

On top of that, as illustrated in the bottom row of Fig. 9, four parallel convolution layers with different kernel sizes are performed on the initially fused features $\mathbf{F}^i_{fus}$, obtaining four scales of features, i.e.,

$$\mathbf{F}^i_m = \mathrm{Conv}_m(\mathbf{F}^i_{fus}, \theta_m), \tag{13}$$

where $\mathrm{Conv}_m(*, \theta_m)$ denotes a convolutional layer with kernel size of m × m ($m = 3, 5, 7, 9$). θ_m refers to their parameters. $\mathbf{F}^i_m$ ($m = 3, 5, 7, 9$) denotes different scales of features in the i-th level.

In addition, considering that different scales of features have certain contributes for final saliency prediction, we propose a scale-aware attention mechanism to adaptively fuse those multi-scale features. Specially, we first feed the multi-scale features into several SE blocks [54], obtaining their scale attention weights $\mathbf{V}^i_m$ ($m = 3, 5, 7, 9$), respectively. Then the softmax function is performed on $\mathbf{V}^i_m$ to obtain the multi-scale weight vector $\mathbf{W}^i$, which reflects the feature importance of different scales. Finally, the multi-scale features are re-weighted by using such importance weight vector, achieving the finally enhanced cross-modal features $\mathbf{F}^i$ in the i-th level. Mathematically, the whole process can be expressed as follows:

$$\mathbf{V}^i_m = \mathrm{SE}(\mathbf{F}^i_m), \tag{14}$$

$$\mathbf{W}^i = \mathrm{Softmax}(\mathrm{Cat}(\mathbf{V}^i_3, \mathbf{V}^i_5, \mathbf{V}^i_7, \mathbf{V}^i_9)), \tag{15}$$

$$\mathbf{F}^i = \mathrm{Cat}(\mathbf{W}^i \otimes \mathrm{Cat}(\mathbf{F}^i_3, \mathbf{F}^i_5, \mathbf{F}^i_7, \mathbf{F}^i_9)). \tag{16}$$

Here, SE $(*)$ denotes the SE block in [54].

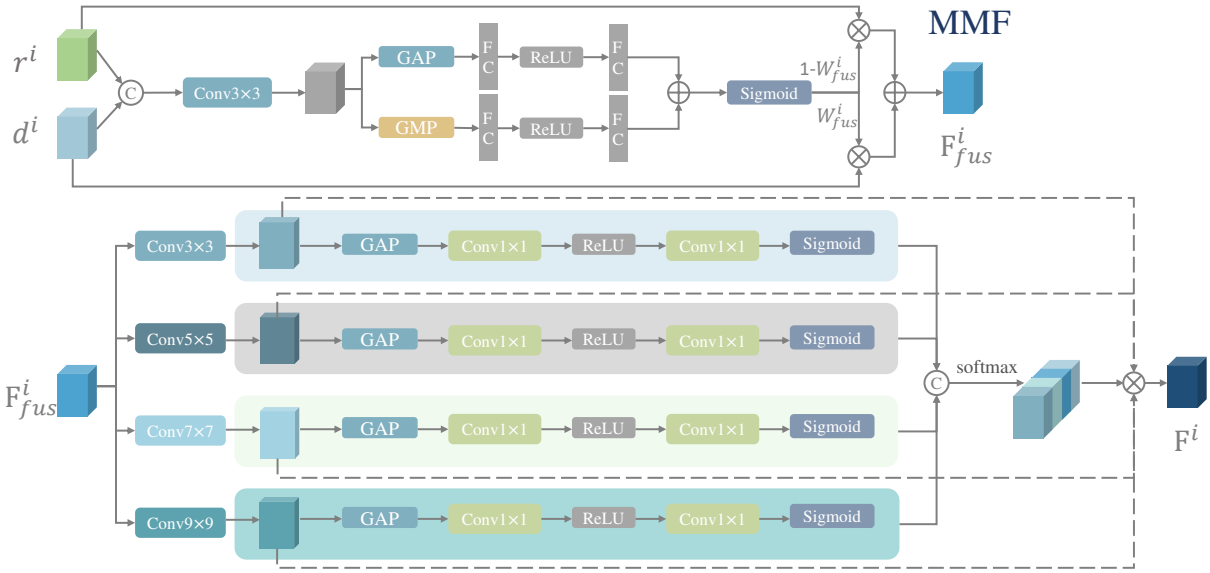


Fig. 9. Architecture of the proposed MMF.

*4) **Loss Function**:*

RCA Loss: The BCE loss [30] and IoU loss [31] are two widely used loss functions in SOD, which provide some pixel-level constraints to force the predicted results close to the GTs. However, they usually ignore the local regional consistency within the saliency maps, resulting in inaccurate saliency detection results. In view of this, we propose a novel auxiliary loss function, i.e., Region Consistency Aware (RCA) loss, where the local regional consistency within the foreground regions and the local regional consistency within the background regions are simultaneously considered.

For that, we first compute the false negative ($\mathbf{FN}$) part in the foreground salient object regions and the false positive ($\mathbf{FP}$) part in the background regions by using Eq. (17) and Eq. (18), respectively. In Eq. (17) and Eq. (18), $\mathbf{S}$ denotes the predicted saliency map and $\mathbf{G}$ denotes its corresponding ground truth. $|\cdot|$ denotes the aboslute value of a number. Given $\mathbf{FN}$ and $\mathbf{FP}$, the proposed RCA loss ℓ_{rca} is computed by using Eq. (19) to enhance the local regional saliency consistency within the foreground salient object regions as well as in the background regions by reducing the saliency differences between each pixel and its surrounding pixels.

$$\mathbf{FN} = |(\mathbf{G} - \mathbf{S}) \otimes \mathbf{G}|, \tag{17}$$

$$\mathbf{FP} = |(\mathbf{S} - \mathbf{G}) \otimes (\mathbf{1} - \mathbf{G})|, \tag{18}$$

$$\begin{aligned}\ell_{rca} = &\tfrac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\mathbf{FN} - \mathrm{Avg}_{\mathrm{spa}}(\mathbf{FN})\right| + \\ &\tfrac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\mathbf{FP} - \mathrm{Avg}_{\mathrm{spa}}(\mathbf{FP})\right|.\end{aligned} \tag{19}$$

In Eq. (19), $\mathrm{Avg}_{\mathrm{spa}}(*)$ denotes the 7×7 average pooling operation along the spatial dimension. H and W denote the height and width of the saliency map, respectively. $\mathbf{1}$ denotes an all-one vector with the same size as $\mathbf{G}$. The proposed RCA loss is beneficial for facilitating the completeness of salient objects and suppressing the introduction of disturbing backgrounds in the saliency map. This will be verified in the experimental part.

Total Loss: Given the proposed RCA loss ℓ_{rca}, a joint loss function ($\mathcal{L}_{joint}$) will be employed in this paper to train our proposed FCFNet, which is defined by

$$\mathcal{L}_{joint}(\mathbf{S}, \mathbf{G}) = \ell_{bce}(\mathbf{S}, \mathbf{G}) + \ell_{iou}(\mathbf{S}, \mathbf{G}) + \lambda * \ell_{rca}(\mathbf{S}, \mathbf{G}), \tag{20}$$

where ℓ_{bce} $(*)$ and $\ell_{iou}(*)$ denote the BCE loss and IoU loss, respectively. The hyper-parameter λ is experimentally set to 15 in the paper.

As shown in Fig. 3 and similar to that in [62], multi-level supervisions are also performed on saliency maps $\mathbf{S}^{(i)}(i = 3, 4, 5)$ in our proposed model, which are deduced from the side-output features of each decoder through a 1×1 convolutional layer and a Sigmoid function. Therefore, the total loss function $\mathcal{L}_{total}$ is expressed by

$$\mathcal{L}_{total} = \sum_{i=3}^{5}(\mathcal{L}_{joint}(\mathrm{Up}(\mathbf{S}^{(i)}), \mathbf{G})). \tag{21}$$

Here, $\mathrm{Up}(*)$ denotes the bilinear upsampling operation.

IV. EXPERIMENTS AND RESULTS

A. Datasets and Evaluation Metrics

1) Datasets: We evaluate the proposed model on six benchmark datasets, including DUT-RGBD [55], NJU2K [63], NLPR [64], STERE [65], LFSD [66] and RGBD135 [13]. DUT-RGBD [55] consists of 1200 paired images captured by Lytro camera in real-life scenes. This dataset is split into 800 training data and 400 testing data. NJU2K [63] consists of 1985 RGB-D stereo images, which are collected from 3D movies, the Internet and Photographs. NLPR [64] consists of 1000 images pairs captured by Kinect under different illumination conditions. STERE [65] consists of 1000 stereoscopic images, where the depth images are estimated from the stereo images. LFSD [66] consists of 100 images with depth

TABLE I
QUANTITATIVE COMPARISONS OF OUR PROPOSED MODEL WITH 21 STATE-OF-THE-ART RGB-D SALIENCY MODELS ON SIX BENCHMARKS DATASETS. THE BEST THREE RESULTS ARE SHOWED IN RED, GREEN AND BLUE COLORS, RESPECTIVELY.

Datasets	Metric	MMCI [18]	TANet [17]	DMRA [55]	CPFP [56]	D3Net [24]	ICNet [19]	A2dele [27]	S2MA [57]	DRLF [58]	FRDT [59]	JL-DCF [60]	SSF [52]	DQSD [28]	CCAFNet [38]	DFMNet [21]	TMFNet [61]	CIRNet [39]	GCENet [47]	CFIDNet [40]	HINet [41]	DCMF [42]	Ours (Ours)
NJU2K	$F_\beta\uparrow$	0.852	0.874	0.886	0.877	0.900	0.891	0.873	0.889	0.883	0.898	0.904	0.896	0.897	0.910	0.913	0.882	0.916	0.915	0.920	0.913	0.888	0.923
	adpF↑	0.812	0.844	0.872	0.837	0.840	0.867	0.874	0.836	0.856	0.878	0.883	0.885	0.840	0.886	0.883	0.874	0.890	0.890	0.889	0.887	0.881	0.891
	$E_\xi\uparrow$	0.915	0.925	0.927	0.923	0.936	0.926	0.916	0.930	0.926	0.933	0.944	0.935	0.936	0.943	0.949	0.910	0.947	0.947	0.913	0.945	0.925	0.953
	adpE↑	0.878	0.896	0.920	0.917	0.892	0.904	0.897	0.896	0.900	0.917	0.912	0.912	0.934	0.919	0.907	0.902	0.918	0.919	0.928	0.922	0.925	0.929
	$S_\alpha\uparrow$	0.858	0.878	0.886	0.878	0.900	0.894	0.869	0.894	0.866	0.898	0.910	0.899	0.899	0.909	0.912	0.910	0.916	0.914	0.914	0.915	0.913	0.918
	MAE ↓	0.079	0.060	0.051	0.053	0.046	0.052	0.051	0.053	0.055	0.048	0.038	0.043	0.050	0.37	0.039	0.041	0.039	0.038	0.038	0.039	0.043	0.034
NLPR	$F_\beta\uparrow$	0.815	0.863	0.880	0.867	0.897	0.908	0.880	0.902	0.870	0.867	0.915	0.896	0.898	0.908	0.910	0.867	0.914	0.907	0.910	0.906	0.875	0.911
	adpF↑	0.747	0.795	0.854	0.821	0.834	0.867	0.870	0.847	0.829	0.863	0.875	0.875	0.842	0.880	0.878	0.854	0.873	0.876	0.875	0.875	0.854	0.879
	$E_\xi\uparrow$	0.913	0.941	0.947	0.932	0.952	0.945	0.953	0.930	0.945	0.963	0.953	0.952	0.956	0.961	0.944	0.963	0.947	0.953	0.950	0.957	0.940	0.960
	adpE↑	0.884	0.916	0.941	0.917	0.932	0.941	0.945	0.937	0.926	0.942	0.952	0.946	0.934	0.948	0.947	0.926	0.949	0.949	0.948	0.948	0.940	0.949
	$S_\alpha\uparrow$	0.856	0.886	0.899	0.888	0.912	0.923	0.896	0.915	0.896	0.914	0.926	0.914	0.916	0.921	0.923	0.921	0.923	0.919	0.922	0.922	0.922	0.924
	MAE ↓	0.059	0.041	0.031	0.036	0.030	0.028	0.028	0.030	0.032	0.029	0.024	0.026	0.029	0.026	0.026	0.027	0.025	0.025	0.026	0.026	0.029	0.024
DUT	$F_\beta\uparrow$	0.767	0.790	0.898	0.779	0.793	0.850	0.892	0.901	0.801	0.903	0.911	0.924	0.827	0.913	0.747	-	0.929	0.923	-	-	0.928	0.931
	adpF↑	0.753	0.778	0.793	0.793	0.755	0.830	0.891	0.886	0.803	0.901	0.882	0.904	0.818	0.903	0.751	-	0.909	0.909	-	-	0.906	0.910
	$E_\xi\uparrow$	0.859	0.861	0.933	0.829	0.829	0.899	0.930	0.937	0.856	0.941	0.943	0.951	0.878	0.943	0.844	-	0.954	0.951	-	-	0.950	0.956
	adpE↑	0.855	0.866	0.868	0.868	0.821	0.897	0.924	0.921	0.872	0.938	0.931	0.946	0.889	0.940	0.850	-	0.945	0.945	-	-	0.943	0.949
	$S_\alpha\uparrow$	0.791	0.808	0.889	0.773	0.773	0.852	0.885	0.903	0.826	0.910	0.881	0.915	0.845	0.903	0.791	-	0.922	0.918	-	-	0.920	0.924
	MAE ↓	0.113	0.093	0.048	0.076	0.098	0.072	0.042	0.043	0.080	0.039	0.055	0.033	0.072	0.037	0.092	-	0.034	0.035	-	-	0.035	0.032
STERE	$F_\beta\uparrow$	0.863	0.861	0.857	0.874	0.891	0.898	0.879	0.882	0.878	0.880	0.898	0.890	0.886	0.887	0.904	-	0.903	0.900	0.903	0.883	0.872	0.906
	adpF↑	0.829	0.835	0.831	0.830	0.859	0.864	0.832	0.856	0.845	0.855	0.862	0.861	0.839	0.860	0.863	-	0.862	0.866	0.864	0.835	0.866	0.868
	$E_\xi\uparrow$	0.927	0.923	0.916	0.925	0.938	0.942	0.928	0.932	0.929	0.927	0.942	0.936	0.935	0.934	0.948	-	0.945	0.940	0.924	0.933	0.930	0.947
	adpE↑	0.901	0.906	0.933	0.918	0.904	0.938	0.913	0.918	0.920	0.899	0.916	0.923	0.908	0.921	0.926	-	0.925	0.925	0.921	0.923	0.926	0.927
	$S_\alpha\uparrow$	0.873	0.871	0.845	0.879	0.899	0.903	0.879	0.890	0.888	0.901	0.900	0.893	0.892	0.892	0.908	-	0.905	0.899	0.901	0.892	0.903	0.906
	MAE ↓	0.068	0.060	0.063	0.051	0.046	0.045	0.045	0.051	0.050	0.043	0.042	0.044	0.051	0.044	0.040	-	0.043	0.040	0.043	0.049	0.043	0.038
LFSD	$F_\beta\uparrow$	0.771	0.796	0.856	0.826	0.810	0.871	0.835	0.835	0.828	0.855	0.867	0.866	0.847	0.832	0.866	0.846	0.873	0.849	0.865	0.847	0.771	0.876
	adpF↑	0.779	0.794	0.849	0.813	0.800	0.861	0.833	0.803	0.823	0.847	0.840	0.840	0.842	0.832	0.863	0.843	0.849	0.850	0.849	0.838	0.768	0.851
	$E_\xi\uparrow$	0.839	0.847	0.900	0.872	0.862	0.903	0.879	0.873	0.873	0.899	0.902	0.900	0.878	0.876	0.902	0.865	0.906	0.883	0.902	0.888	0.842	0.913
	adpE↑	0.840	0.845	0.899	0.867	0.833	0.891	0.869	0.863	0.856	0.889	0.882	0.895	0.882	0.871	0.891	0.851	0.899	0.885	0.891	0.886	0.822	0.902
	$S_\alpha\uparrow$	0.787	0.801	0.847	0.828	0.825	0.868	0.836	0.837	0.834	0.857	0.833	0.859	0.851	0.826	0.870	0.849	0.869	0.846	0.870	0.852	0.870	0.875
	MAE ↓	0.132	0.111	0.075	0.088	0.095	0.071	0.074	0.094	0.089	0.073	0.084	0.066	0.085	0.087	0.068	0.084	0.069	0.079	0.070	0.076	0.068	0.061
RGBD135	$F_\beta\uparrow$	0.822	0.827	0.888	0.846	0.885	0.913	0.867	0.935	0.869	0.868	0.926	0.883	0.927	0.937	0.932	0.892	-	0.925	0.934	-	0.901	0.939
	adpF↑	0.750	0.794	0.857	0.818	0.874	0.877	0.868	0.893	0.868	0.861	0.886	0.882	0.881	0.904	0.895	0.876	-	0.904	0.898	-	0.896	0.908
	$E_\xi\uparrow$	0.928	0.910	0.945	0.923	0.946	0.960	0.923	0.973	0.940	0.942	0.969	0.941	0.973	0.977	0.973	0.968	-	0.966	0.943	-	0.967	0.980
	adpE↑	0.899	0.919	0.942	0.927	0.956	0.945	0.919	0.971	0.932	0.923	0.966	0.946	0.967	0.975	0.971	0.965	-	0.966	0.940	-	0.967	0.978
	$S_\alpha\uparrow$	0.848	0.858	0.901	0.872	0.898	0.920	0.885	0.937	0.895	0.902	0.924	0.905	0.935	0.938	0.938	0.936	-	0.926	0.917	-	0.881	0.939
	MAE ↓	0.065	0.046	0.029	0.038	0.031	0.027	0.028	0.021	0.030	0.028	0.021	0.025	0.021	0.018	0.019	0.021	-	0.018	0.023	-	0.023	0.017

information and human-labeled ground-truths. RGBD135 [13] consists of 135 RGB-D images from seven indoor scenarios.

2) Evaluation Metrics: To quantitatively evaluate the effectiveness of our proposed model, we use some standard metrics, including Precision-Recall (PR) [67] [68], Mean Absolute Error (MAE) [69], maximum F-measure (F_β) [67], adptive F-measure (adpF) [67], S-measure (S_α) [70], maximum E-measure (E_ξ) [71] and adptive E-measure (adpE) [71]. Larger values of F_β, S_α, E_ξ, adpF, adpE and smaller values of MAE are more desirable for an SOD method.

B. Implementation Details

We implement the proposed model by PyTorch [72] with an NVIDIA 1080Ti GPU. As the same splitting way in [27] [55], we select 800 samples from DUT-RGBD [55], 1485 samples from NJU2K [63] and 700 samples from NLPR [64] for training. The remaining images in the three datasets and those images in the remaining two datasets are all for testing to verify the performance of different models. At the training and testing phase, all the input images are resized to 224×224. Random horizontal flipping and random vertical flipping are adopted for data augmentation. The image generation network is trained with Adam optimizer [73], where the batch size, learning rate and weight decay are set to 4, 1e-4 and 4e-5, respectively. The FCFNet is trained with the Stochastic Gradient Descent (SGD) optimizer [74]. Here, the batch size, momentum and weight decay of the SGD optimizer are set to 5, 0.9 and 5e-4, respectively. Meanwhile, the learning rate is set to 1e-4, which is divided by 10 after 35 epochs. The maximum epochs of our network is 70.

C. Comparison with State-of-the-Art Models

We compare our proposed model with 21 state-of-the-art RGB-D SOD methods, including MMCI [18], TANet [17], DMRA [55], CPFP [56], D3Net [24], ICNet [19], A2dele [27], S2MA [57], DRLF [58], FRDT [59], JL-DCF [60], SSF [52], CMWNet [75], DQSD [28], DFMNet [21], CCAFNet [38], CIRNet [39],GCENet[47], CFIDNet [40], HINet [41] and DCMF [42]. For fair comparisons, the saliency maps of these SOTA models are obtained from their authors or the deployment codes provided by their authors.

Some visualization results are illustrated in Fig. 10. As shown in the first two rows of Fig. 10, all of the methods mentioned here perform well for those images with simple scenes. However, as shown in the 3^{rd}, 4^{th}, 5^{th} and 6^{th} rows of Fig. 10, when the depth images have low visual qualities or have some inconsistent foregrounds with their corresponding RGB images, those SOTA methods cannot obtain desirable detection results. For example, some salient objects are not completely detected, or parts of the backgrounds are not well suppressed during the saliency detection. Moreover, as shown in the last two rows of Fig. 10, when the scenarios have multiple objects, most SOTA models fail to achieve complete detection for all salient objects. While our method can still effectively detect the salient objects for these challenging scenes. This may be due to the fact that such undesirable information in the original low-quality or foreground-inconsistent depth images are well calibrated by using those generated pseudo depth images. As well, the multi-scale context information may be effectively exploited for saliency detection in our method because of the proposed MMF module. As shown in Table I and Fig. 11, it can be easily seen that our proposed model outperforms those SOTA methods in terms of the five metrics, i.e., F_β, S_α, E_ξ, MAE and PR curves on the NJU2K [63], DUT-RGBD [55] and LFSD [66] datasets. For NLPR [64] and STERE [65], our proposed model also achieves competitive results with these SOTA methods.

D. Ablation Study

Actually, our proposed method consists of two stages, i.e., an image generation stage and a saliency reasoning stage.

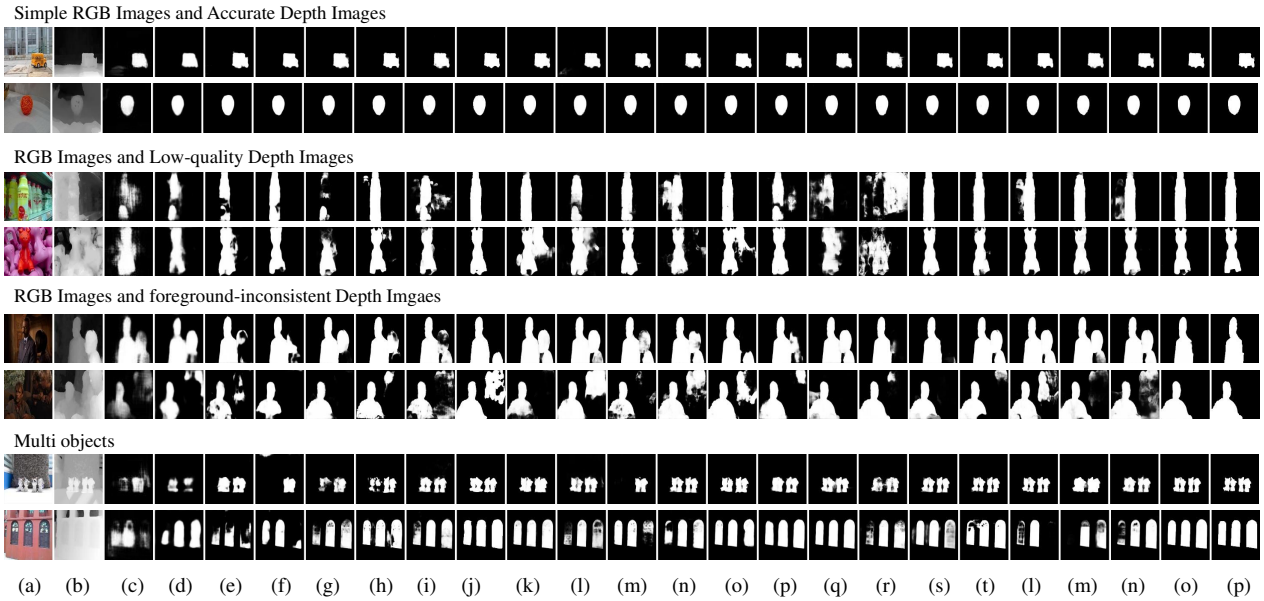
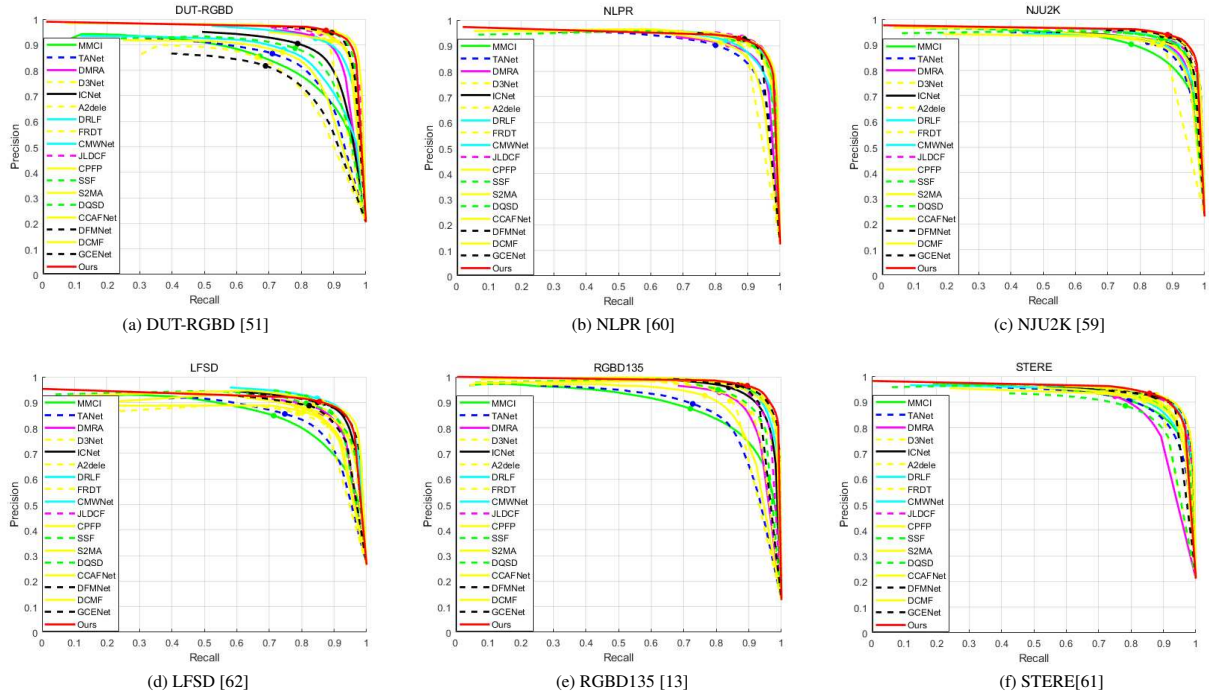


Fig. 10. Visual comparisons of different methods. (a) RGB images; (b) Depth images; (c) MMCI [18]; (d) TANet [17]; (e) CPFP [56]; (f) DMRA [55]; (g) D3Net [24]; (h) SSF [52]; (i) ICNet [19]; (j) A2dele [27]; (k) S2MA [57]; (l) DRLF [58]; (m) FRDT [59]; (n) CMWNet [75]; (o) CCAFNet [38]; (p) JL-DCF [60]; (q) DQSD [28]; (r) DFMNet [21]; (s) CIRNet [39]; (t)GCENet [47]; (l) CFIDNet [40]; (m) HINet [41]; (n) DCMF [42]; (o) FCFNet (Ours); (p) GTs.



(a) DUT-RGBD [51] (b) NLPR [60] (c) NJU2K [59]

(d) LFSD [62] (e) RGBD135 [13] (f) STERE[61]

Fig. 11. Quantitative comparisons of our proposed method with other methods on six benchmark datasets.

Thus, we conduct two sets of ablation experiments on the NJU2K dataset [63] to verify the effectiveness of each component contained in each stage, respectively. First, we verify the effectiveness of the TSS strategy proposed in the image generation stage. Then, we verify the effectiveness of each component in FCFNet proposed in the saliency reasoning stage.

(1) Validity of TSS strategy

In order to demonstrate the effectiveness of the proposed TSS strategy, we first construct a baseline model, denoted as 'B', where the DFC and SFI modules, together with the RCA loss, are removed from FCFNet. As well, MMF is replaced with some element-wise addition operations. On top of that, we construct four versions of our proposed TSS strategy by using different sample selection ways, i.e., TSS (w/o 'step1' and 'step2'), TSS (w/o 'step1'), TSS (w/o 'step2') and TSS. In TSS (w/o 'step1' and 'step2'), 'step1' and 'step2'

TABLE II
QUANTITATIVE EVALUATION OF ABLATION STUDIES OF TSS ON NJU2K DATASET.

Methods	MAE	F_β	S_α	E_ζ
B	0.042	0.907	0.906	0.941
B+TSS (w/o 'step1' and 'step2')	0.042	0.906	0.905	0.942
B+TSS (w/o 'step1')	0.040	0.905	0.907	0.941
B+TSS (w/o 'step2')	0.040	0.907	0.907	0.941
B+TSS	0.039	0.913	0.909	0.945

are simultaneously removed from our proposed TSS strategy, which means that we use all original depth images as the supervision information for the image generation network. In TSS (w/o 'step1') and TSS (w/o 'step2'), 'step1' and 'step2' are removed from our proposed TSS strategy, respectively.

The corresponding quantitative results of different versions are provided in Table II. It can be seen that the method of using all original depth images as the supervision information even deteriorates the performance of 'B'. This is due to the fact that those original low-quality depth images will contaminate the qualities of the generated pseudo depth images, thus leading to unsatisfactory results. Compared to utilizing all original depth images as supervision information, the performance of 'B' is enhanced after employing the proposed TSS, which indicates the validity of TSS. Meanwhile, we can also observe that the performance of 'B+TSS' is degraded when the 'step1' or 'step2' is removed, which indicates that each selection step is beneficial for the generation of pseudo depth images.

Furthermore, as shown in Fig. 12, we also provide some visual comparisons of the generated pseudo depth images to further verify the validity of TSS. As shown in Fig. 12(c), since those original low-quality and foreground-inconsistent depth images are utilized to supervise the image generation network, the pseudo depth images generated by TSS (w/o 'step1' and 'step2') still have low visual qualities or have foreground inconsistency with their corresponding RGB images. Differently, as shown in the first two rows of Fig. 12(e), the visual qualities of the pseudo depth images generated by TSS (w/o 'step2') are significantly improved by removing those low-quality depth images as supervision information. However, as shown in the last two rows of Fig. 12(e), such method may fail for those depth images that have inconsistent foregrounds with their corresponding RGB images. Similar phenomena are reported for TSS (w/o 'step1'). Specifically, the pseudo depth images generated by TSS (w/o 'step1') have consistent foregrounds with their corresponding RGB images but their visual qualities still have large room for improvement. By comparing Fig. 12(d), Fig. 12(e) and Fig. 12(f), it is easily observed that better pseudo depth images can be obtained by using the TSS strategy than by other versions. This indicates that our proposed TSS strategy can better select those high-quality and foreground-consistent depth images for supervising the image generation network than TSS (w/o 'step1') and TSS (w/o 'step2') do, thus generating better pseudo depth images for calibrating the original depth information.

(2) Validity of each component in FCFNet

Specifically, the following five models are mainly involved in our ablation study to demonstrate the effectiveness of each component in FCFNet:

- **FCFNet w/o DFC**: FCFNet without the DFC module.
- **FCFNet w/o SFI**: FCFNet without the SFI module.
- **FCFNet w/o MMF**: Replacing the MMF module with element-wise addition in FCFNet.
- **FCFNet w/o RCA**: FCFNet without the RCA loss.
- **FCFNet**: Our proposed model.

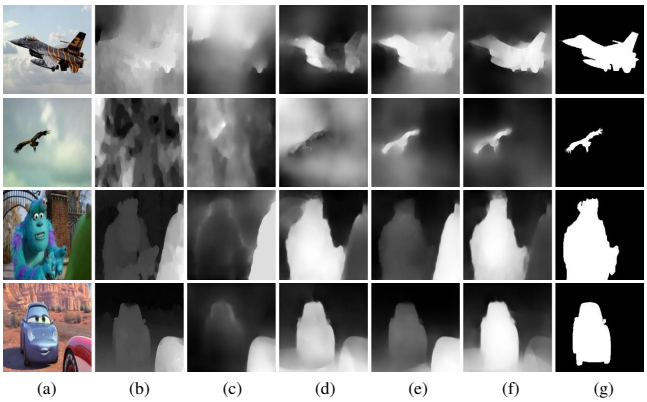


Fig. 12. Visualization of our proposed TSS strategy with different versions. (a) RGB images; (b) Depth images; (c) Generated pseudo depth images by using TSS (w/o 'step1' and 'step2'); (d) Generated pseudo depth images by using TSS (w/o 'step1'); (e) Generated pseudo depth images by using TSS (w/o 'step2'); (f) Generated pseudo depth images by using TSS; (g) GTs for saliency maps.

TABLE III
QUANTITATIVE EVALUATION OF ABLATION STUDIES OF FCFNET ON NJU2K DATASET.

Methods	MAE	F_β	S_α	E_ζ
FCFNet w/o SFI	0.036	0.921	0.915	0.950
FCFNet w/o DFC	0.036	0.923	0.917	0.951
FCFNet w/o MMF	0.037	0.911	0.911	0.945
FCFNet w/o RCA	0.037	0.920	0.913	0.948
FCFNet	0.034	0.924	0.918	0.952

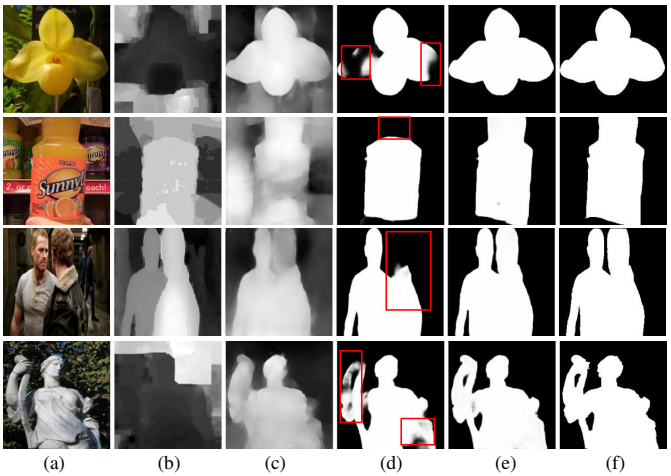


Fig. 13. Visual comparisons between FCFNet w/o DFC and FCFNet. (a) RGB images; (b) Depth images; (c) Generated pseudo depth images; (d) Saliency maps obtained by FCFNet w/o DFC; (e) Saliency maps obtained by FCFNet; (f) GTs. The red boxes indicate the obvious differences.

Validity of DFC: In FCFNet w/o DFC, the features ex-

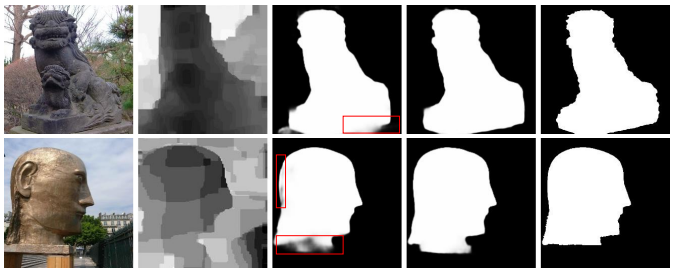


Fig. 14. Visual comparisons between FCFNet w/o SFI and FCFNet. (a) RGB images; (b) Depth images; (c) Saliency maps obtained by FCFNet w/o SFI; (d) Saliency maps obtained by FCFNet; (e) GTs. The red boxes indicate the obvious differences.

TABLE IV
QUANTITATIVE EVALUATION OF ABLATION STUDIES OF MMF ON NJU2K DATASET.

Methods	MAE	F_β	S_α	E_ζ
MMF w/o MA	0.036	0.915	0.911	0.949
MMF w/o MA + ASPP	0.036	0.919	0.916	0.951
MMF w/o MA + DASPP	0.038	0.916	0.913	0.948
MMF	0.034	0.923	0.918	0.953

tracted from the original depth images will be directly fed into the subsequent fusion module to achieve saliency prediction. As shown in Table III, we can observe that the performance of FCFNet w/o DFC is worse than the proposed FCFNet. This indicates that the effectiveness of our DFC module for the calibration of original depth features. Furthermore, as shown in Fig. 13, it can be observed that the influence of those unreliable information from the original depth image can be effectively reduced with the introduction of DFC.

Validity of SFI: As reported in Table III, SFI also promotes the performance of saliency detection by injecting the detailed information from the shallower levels into the middle level of features. Intuitively, as shown in Fig. 14, the detected salient objects achieve sharper boundaries by virtue of our proposed SFI module. This indicates that SFI can effectively exploit the detailed information contained in the shallower-level features for better saliency prediction results.

Validity of MMF: It can be seen from the quantitative results in Table III that replacing our proposed MMF module with element-wise addition operation will decrease the performance of FCFNet. This demonstrates that MMF can facilitate the cross-modal and multi-scale complementary information exploration for SOD. Moreover, we also provide some visual comparisons to verify the effectiveness of our proposed MMF module in Fig. 15, which demonstrates that multiple salient objects can be simultaneously detected by using our proposed MMF module.

Especially, in order to further demonstrate the validity of MMF for multi-scale information exploitation, we construct different versions of our proposed MMF module by keeping CF unchanged and replacing MA with different modules, which are denoted as MMF w/o MA, MMF w/o MA+ASPP and MMF w/o MA+DASPP, respectively. In MMF w/o MA, MA is directly removed from our proposed MMF, which denotes that we do not further exploit the multi-scale contextual

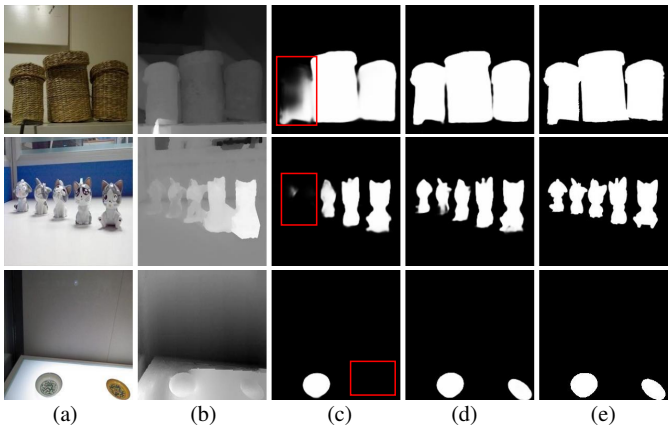


Fig. 15. Visual comparisons between FCFNet w/o MMF and FCFNet. (a) RGB images; (b) Depth images; (c) Saliency maps obtained by FCFNet w/o MMF; (d) Saliency maps obtained by FCFNet; (e) GTs. The red boxes indicate the obvious differences.

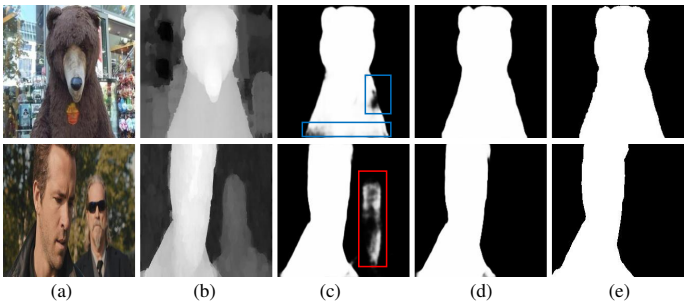


Fig. 16. Visual comparisons between FCFNet w/o RCA and FCFNet. (a) RGB images; (b) Depth images; (c) Saliency maps obtained by FCFNet w/o RCA; (d) Saliency maps obtained by FCFNet; (e) GTs. The red boxes indicate the false positive regions and the blue boxes indicate the false negative regions.

information for SOD. In MMF w/o MA+ASPP, the ASPP module is added to MMF w/o MA, where the dilation rates are set to 6/12/18/24 according to the literature [76]. In MMF w/o MA+DASPP, the DASPP module is added to MMF w/o MA, where the dilation rates are set to 3/6/12/18/24 according to the literature [77]. As shown in Table IV, our proposed MMF obtains better performance than other modules. This indicates that MMF can more effectively capture the multi-scale information from the cross-modal RGB-D features via the proposed MA sub-module for SOD.

Validity of RCA: As shown in Table III, the performance of SOD is significantly degraded after removing the RCA loss. This can be also verified in Fig. 16. As shown in the 1^{st} row of Fig. 16, incomplete foregrounds are easily obtained if the RCA loss is removed from the joint loss function defined by Eq. (20). Similarly, as shown in the 2^{nd} row of Fig. 16, some backgrounds are also mistakenly detected as salient ones without using the proposed RCA loss function in FCFNet. This indicates that the local regional saliency consistency within the foreground salient object regions as well as within the background regions is beneficial for the accurate segmentation of salient objects, thus achieving better saliency results by using our proposed RCA loss.

V. Conclusion

In this paper, a two-stage RGB-D salient object detection model has been presented, which is composed of an image generation stage and a saliency reasoning stage. In the image generation stage, owing to the proposed TSS strategy, high-quality and foreground-consistent pseudo depth images can be generated from the input RGB images. In the saliency reasoning stage, the original depth features are first calibrated by using the generated pseudo depth images via the proposed DFC module and then fused with the RGB features for saliency prediction via the proposed SFI and MMF modules. By virtue of the proposed calibration-then-fusion strategy, the influence of such low-quality depth images as well as that of those foreground-inconsistent depth images on the saliency prediction can be greatly reduced. Moreover, thanks to the proposed RCA auxiliary loss function in the saliency reasoning stage, where the local regional saliency consistency within the foreground salient object regions and that within the background regions are both considered, more complete salient objects and less disturbing backgrounds can be obtained in the final saliency maps. Experimental results on six benchmark datasets demonstrate the validity of our proposed RGB-D SOD model, especially when the depth images have low visual qualities, or have some inconsistent foregrounds with their corresponding RGB images in the scenes.

Acknowledgment

This work is supported by the Research Foundation of the Key Laboratory of Spaceborne Information Intelligent Interpretation No. 2022-ZZKY-JJ-09-01, by the Shaanxi Innovation Team Project 2018TD-012, and by the National Natural Science Foundation of China under Grant No.61773301.

References

[1] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 769–779, 2013.

[2] W. Wang, J. Shen, H. Sun, and L. Shao, "Video co-saliency guided co-segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1727–1736, 2017.

[3] J. Lei, L. Niu, H. Fu, B. Peng, Q. Huang, and C. Hou, "Person re-identification by semantic region representation and topology constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2453–2466, 2018.

[4] R. Yao, G. Lin, C. Shen, Y. Zhang, and Q. Shi, "Semantics-aware visual object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 6, pp. 1687–1700, 2018.

[5] Y. Niu, H. Zhang, W. Guo, and R. Ji, "Image quality assessment for color correction based on color contrast similarity and color value difference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 4, pp. 849–862, 2016.

[6] W.-D. Jin, J. Xu, M.-M. Cheng, Y. Zhang, and W. Guo, "Icnet: Intra-saliency correlation network for co-saliency detection," *Advances in Neural Information Processing Systems*, 2020.

[7] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3917–3926.

[8] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5968–5977.

[9] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7234–7243.

[10] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328.

[11] J. Li, Z. Pan, Q. Liu, and Z. Wang, "Stacked u-shape network with channel-wise attention for salient object detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 1397–1409, 2020.

[12] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," *arXiv preprint arXiv:1511.03643*, 2015.

[13] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proceedings of international conference on internet multimedia computing and service*, 2014, pp. 23–27.

[14] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang, "Exploiting global priors for rgb-d saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 25–32.

[15] X. Xiao, Y. Zhou, and Y.-J. Gong, "Rgb-d saliency detection with pseudo depth," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2126–2139, 2018.

[16] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3051–3060.

[17] H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[18] 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 Recognition*, vol. 86, pp. 376–385, 2019.

[19] G. Li, Z. Liu, and H. Ling, "Icnet: Information conversion network for rgb-d based salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 4873–4884, 2020.

[20] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, 2020.

[21] W. Zhang, G.-P. Ji, Z. Wang, K. Fu, and Q. Zhao, "Depth quality-inspired feature manipulation for efficient rgb-d salient object detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 731–740.

[22] X. Wang, S. Li, C. Chen, A. Hao, and H. Qin, "Depth quality-aware selective saliency fusion for rgb-d image salient object detection," *Neurocomputing*, vol. 432, pp. 44–56, 2021.

[23] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for rgb-d salient object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 2020, pp. 235–252.

[24] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on neural networks and learning systems*, vol. 32, no. 5, pp. 2075–2089, 2020.

[25] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "Cdnet: Complementary depth network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3376–3390, 2021.

[26] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4296–4307, 2020.

[27] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9060–9069.

[28] C. Chen, J. Wei, C. Peng, and H. Qin, "Depth-quality-aware salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 2350–2363, 2021.

[29] X. Zhang, Y. Xu, T. Wang, and T. Liao, "Multi-prior driven network for rgb-d salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[30] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[32] B. Wang, Q. Chen, M. Zhou, Z. Zhang, X. Jin, and K. Gai, "Progressive

feature polishing network for salient object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 128–12 135.

[33] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7479–7489.

[34] Y. Kong, M. Feng, X. Li, H. Lu, X. Liu, and B. Yin, "Spatial context-aware network for salient object detection," *Pattern Recognition*, vol. 114, p. 107867, 2021.

[35] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3203–3212.

[36] H. Mei, Y. Liu, Z. Wei, D. Zhou, X. Wei, Q. Zhang, and X. Yang, "Exploring dense context for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1378–1389, 2021.

[37] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3528–3542, 2021.

[38] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "Ccafnet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in rgb-d images," *IEEE Transactions on Multimedia*, 2021.

[39] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, "Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 6800–6815, 2022.

[40] T. Chen, X. Hu, J. Xiao, G. Zhang, and S. Wang, "Cfidnet: cascaded feature interaction decoder for rgb-d salient object detection," *Neural Computing and Applications*, vol. 34, no. 10, pp. 7547–7563, 2022.

[41] H. Bi, R. Wu, Z. Liu, H. Zhu, C. Zhang, and T.-Z. Xiang, "Cross-modal hierarchical interaction network for rgb-d salient object detection," *Pattern Recognition*, vol. 136, p. 109194, 2023.

[42] F. Wang, J. Pan, S. Xu, and J. Tang, "Learning discriminative cross-modality features for rgb-d saliency detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 1285–1297, 2022.

[43] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2091–2106, 2021.

[44] X. Jin, K. Yi, and J. Xu, "Moadnet: Mobile asymmetric dual-stream networks for real-time and lightweight rgb-d salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7632–7645, 2022.

[45] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency." Georgia Institute of Technology, 2013.

[46] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "Rgbd salient object detection via disentangled cross-modal fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 8407–8416, 2020.

[47] C. Xia, S. Duan, X. Gao, Y. Sun, R. Huang, and B. Ge, "Gcenet: Global contextual exploration network for rgb-d salient object detection," *Journal of Visual Communication and Image Representation*, vol. 89, p. 103680, 2022.

[48] Y. Yang, Q. Qin, Y. Luo, Y. Liu, Q. Zhang, and J. Han, "Bi-directional progressive guidance network for rgb-d salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[49] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2d face recognition via discriminative face depth estimation," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 140–147.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[52] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3472–3481.

[53] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[54] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[55] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7254–7263.

[56] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936.

[57] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765.

[58] X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, and H. Qin, "Data-level recombination and lightweight fusion scheme for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 458–471, 2020.

[59] M. Zhang, Y. Zhang, Y. Piao, B. Hu, and H. Lu, "Feature reintegration over differential treatment: A top-down and adaptive fusion network for rgb-d salient object detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 4107–4115.

[60] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3052–3062.

[61] W. Zhou, S. Pan, J. Lei, and L. Yu, "Tmfnet: Three-input multilevel fusion network for detecting salient objects in rgb-d images," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[62] N. Huang, Y. Liu, Q. Zhang, and J. Han, "Joint cross-modal and unimodal features for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 2428–2441, 2020.

[63] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 1115–1119.

[64] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *European conference on computer vision*. Springer, 2014, pp. 92–109.

[65] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 454–461.

[66] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813.

[67] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1597–1604.

[68] B. Magnier, H. Abdulrahman, and P. Montesinos, "A review of supervised edge detection evaluation methods and an objective comparison of filtering gradient computations using hysteresis thresholds," *Journal of Imaging*, vol. 4, no. 6, p. 74, 2018.

[69] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5706–5722, 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 *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4548–4557.

[71] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.

[72] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[74] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[75] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for rgb-d salient object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 665–681.

[76] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[77] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic

segmentation in street scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3684–3692.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001,2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. His current research interests include image processing, pattern recognition.

Qi Qin received the B.S. degree from Xidian University, Xi'an ,China, in 2018. she is currently pursing M.S. degree in School of Mechano-Electronic Engineering, Xidian University,China. Her current research interests include multimodal image processing and deep learning.

Yang Yang received his B. S. degree from Chang'an University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree in School of Mechano-Electronic Engineering, Xidian University, China. His current research interests include multi-modal image processing and deep learning.

Qiang Jiao received his B.S. and Ph.D. degrees from the Nanjing University of Science and Technology, China in 2010 and 2017, respectively. He is currently working at the School of Mechano-electronic Engineering, Xidian University, China. From March 2014 to March 2015, he was a visiting scholar at the Nanyang Technological University, Singapore. He has also held research positions at the City University of Hong Kong. His current research interests include reinforcement learning and computer vision.

Jungong Han is currently a Chair Professor and the Director of the Department of Computer Science, Aberystwyth University, Aberystwyth, U.K. He also holds an Honorary Professorship with the University of Warwick, Coventry, U.K. His research interests include computer vision, artificial intelligence, and machine learning.