Hierarchical Alternate Interaction Network for RGB-D Salient Object Detection

Gongyang Li, *Member, IEEE*, Zhi Liu, *Senior Member, IEEE*, Minyu Chen, Zhen Bai, Weisi Lin, *Fellow, IEEE*, and Haibin Ling

***Abstract*— Existing RGB-D Salient Object Detection (SOD) methods take advantage of depth cues to improve the detection accuracy, while pay insufficient attention to the quality of depth information. In practice, a depth map is often with uneven quality and sometimes suffers from distractors, due to various factors in the acquisition procedure. In this article, to mitigate distractors in depth maps and highlight salient objects in RGB images, we propose a Hierarchical Alternate Interactions Network (HAINet) for RGB-D SOD. Specifically, HAINet consists of three key stages: feature encoding, cross-modal alternate interaction, and saliency reasoning. The main innovation in HAINet is the Hierarchical Alternate Interaction Module (HAIM), which plays a key role in the second stage for cross-modal feature interaction. HAIM first uses RGB features to filter distractors in depth features, and then the purified depth features are exploited to enhance RGB features in turn. The alternate RGB-depth-RGB interaction proceeds in a hierarchical manner, which progressively integrates local and global contexts within a single feature scale. In addition, we adopt a hybrid loss function to facilitate the training of HAINet. Extensive experiments on seven datasets demonstrate that our HAINet not only achieves competitive performance as compared with 19 relevant state-of-the-art methods, but also reaches a real-time processing speed of 43 *fps* on a single NVIDIA Titan X GPU. The code and results of our method are available at https://github.com/MathLee/HAINet.**

***Index Terms*— RGB-D salient object detection, hierarchical structure, alternate interaction.**

I. Introduction

SALIENT object detection (SOD) is an essential and important task in computer vision. The goal of SOD is to detect and highlight the most salient objects in visual input, such as color images, RGB-D images and videos. It has been applied to many other computer vision tasks, such as visual tracking [1], image captioning [2], weakly supervised learning [3], object segmentation [4], [5], *etc*. Several surveys on color image SOD [6]–[8], RGB-D SOD [9], [10] and video SOD [11], [12] summarize recent developments of SOD in detail. Since the distance-to-camera cues of depth maps naturally supplement appearance information from RGB images for SOD, RGB-D SOD has recently attracted increasing amount of research attention, especially considering the popularity of affordable RGB-D sensors. Numerous RGB-D SOD methods [13]–[41] have been proposed for this purpose and substantial advancements have been achieved.

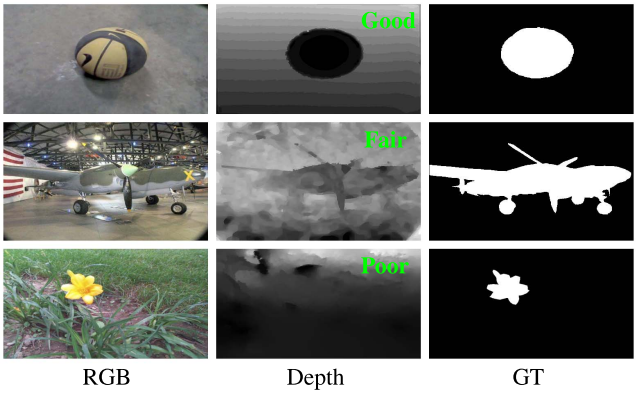


Fig. 1. Examples of depth maps with good, fair and poor quality (from top to bottom), from the DES [42], NJU2K [43] and NLPR [44] datasets, respectively. GT represents ground truth.

Recently, deep learning has shown tremendous capabilities in various computer vision tasks, especially the pixel-level prediction task. The performance of the latest deep learning-based RGB-D SOD methods [16]–[20] surpasses that of traditional RGB-D SOD methods [45]–[49] by a large margin. Traditional RGB-D SOD methods usually regard the depth map as a prior and extract the distance information by hand-crafted operators to assist SOD. Deep learning-based RGB-D SOD methods, which are usually data-driven, change the mindset and adaptively explore complementary information in depth maps in an end-to-end fashion.

However, in both traditional and deep learning-based methods, most works often ignore the quality of depth maps. As a result, the distractors in depth maps often bring troubles to RGB-D SOD. In Fig. 1, we show some examples of depth

Manuscript received September 24, 2020; revised January 25, 2021 and February 24, 2021; accepted February 24, 2021. Date of publication March 5, 2021; date of current version March 11, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61771301 and in part by the Singapore Ministry of Education Tier-2 Fund MOE2016-T2-2-057(S). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Weiyao Lin. *(Corresponding author: Zhi Liu.)*

Gongyang Li, Zhi Liu, Minyu Chen, and Zhen Bai are with the 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; luminouscmy@163.com; bz536476@163.com).

Weisi Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

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.2021.3062689

maps with different qualities, *i.e.*, good, fair and poor, from existing datasets. Obviously, a good depth map shows clear boundary and accurate localization of the object. A fair one can basically locate the object, but has fuzzy boundary. A poor one fails to provide reliable information, may even has a negative impact on RGB features in the cross-modal feature interaction, and consequently hurts the SOD performance. Only a few previous studies have considered the distractors in depth maps and proposed several anti-jamming modules. For example, contrast enhanced net [50] enhances the contrast of depth maps; two-phase depth estimation [21] enlarges the depth differences of depth maps; depth depurator unit [9] judges the quality of depth maps; and cross-modal attention unit [26] selects useful regions from depth maps.

Inspired by the above observation, in this article, we focus on reducing the negative effects of inaccurate depth maps and exploring efficient cross-modal interactions. For this purpose, we propose a novel *Hierarchical Alternate Interaction Module* (HAIM), which is based on the Alternate Interaction Unit (AIU). AIU follows the RGB-depth-RGB flow for purifying depth features (RGB-depth modulation) and enhancing RGB features (depth-RGB feedback). To further increase the efficiency of AIU within a single feature scale, we extend AIU to a hierarchical version. Furthermore, to select key features from multiple AIUs, we perform feature re-weighting via channel attention mechanism. In summary, our module has three advantages different from [9], [21], [26], [50]: the flexible modulation-feedback mechanism (AIU), the ability to capture local and global contexts (hierarchical structure), and the adaptive feature re-weighting operation (channel attention).

To use HAIM for RGB-D SOD, we embed HAIM into an encoding-reasoning architecture at multiple feature scales, and propose a simple yet effective *Hierarchical Alternate Interaction Network* (HAINet). Being a feature fusion-based method, HAINet works in three sequential stages: feature encoding, cross-modal alternate interaction, and saliency reasoning. Feature encoding is responsible for extracting RGB and depth features using a two-stream backbone. Cross-modal alternate interaction, as its name suggests, is in charge of interaction and communication of cross-modal features in HAIM. Finally, the reasoning stage is responsible for the emergence of specific salient objects based on the output of the second stage. In this way, the proposed HAINet fully exploits the potential of HAIM with the effective encoding-reasoning architecture, and overcomes the uncertainty in depth map quality and generates accurate saliency maps.

Our major contributions are summarized as follows:

- We propose a novel Hierarchical Alternate Interaction Module (HAIM), following the RGB-depth-RGB flow paradigm, to effectively enhance the cross-modal interactions between RGB and depth features. In HAIM, the hierarchical structure provides contextual features; the alternate interaction unit performs modulation-feedback mechanism progressively; and the feature re-weighting aims at keeping the most valuable information.
- We apply HAIM to an encoding-reasoning architecture at multiple feature scales, and propose a simple yet effective Hierarchical Alternate Interaction Network (HAINet) for RGB-D SOD, which successfully mitigates distractors in depth maps and accurately highlights specific salient objects in RGB images.
- Comprehensive experiments on seven popular benchmark datasets under five widely used evaluation metrics demonstrate that the proposed HAINet is very competitive to 19 state-of-the-art RGB-D SOD methods, and runs at 43 *fps* on a single NVIDIA Titan X GPU.

The rest of this article is organized as follows: we review the related work of RGB-D SOD in Sec. II; we elaborate the proposed HAINet in Sec. III; we conduct extensive experiments to confirm the superiority and effectiveness of our HAINet in Sec. IV; and we give the conclusion in Sec. V.

II. Related Work

In this section, we briefly review existing representative works on RGB-D salient object detection, including traditional (non-deep learning) methods and deep learning-based methods, and salient object detection on multiple visual media, including color image, light field image and RGB-T image.

A. Traditional RGB-D Salient Object Detection

Many efforts have been spent on investigating traditional (non-deep learning) methods for RGB-D SOD. Their aims are to mine the distance cues of depth maps based on various hand-crafted features. Niu *et al*. [51] tried stereoscopic saliency detection for the first time and proposed the STEREO dataset. They computed the disparity map between left and right views of a stereoscopic image to extract depth cues. Following [51], some studies [44]–[46] also obtain depth cues from stereoscopic images, and used their contrast information. Different from above methods, Li *et al*. [48] utilized the distance clues contained in light field images. Subsequently, Cheng *et al*. [42] and Peng *et al*. [44] built the DES dataset and NLPR dataset, respectively, for RGB-D SOD. Both datasets directly provide depth maps collected by the Kinect device.

The emergence of these datasets largely stimulates the study in RGB-D SOD. Ren *et al*. [52] captured the depth prior and regarded it as one of the global prior. In [53], Feng *et al*. re-weighted the local background enclosure feature with depth and spatial prior. Guo *et al*. [54] introduced the cellular automata to iteratively propagate the initial saliency map to generate the final one. Wang *et al*. [55] obtained depth saliency, depth bias and 3D prior from the depth map, and employed minimum barrier distance for saliency optimization. In [13], Song *et al*. applied the multi-scale pre-segmentation and multi-level saliency computation to color image and depth map, and then fused them. Based on the existing color image SOD methods, Cong *et al*. [22] proposed a transformation strategy to incorporate depth map into those methods to boost the performance of RGB-D SOD. Contrary to [22], Xiao *et al*. [15] introduced the pseudo depth to color image SOD, which achieves promising performance on color image SOD.

Although these traditional methods have achieved promising performance, the distractors in depth maps and the

hand-crafted features limit their generalization. We pay attention to the quality of depth maps and aim to mitigate negative impact of distractors. Concretely, we adapt the Convolutional Neural Network (CNN) to learn unique characteristics of the data through specific encoding-reasoning architecture and HAIMs.

B. Deep Learning-Based RGB-D Salient Object Detection

In recent years, deep learning-based methods have achieved significant progress in RGB-D SOD. Among the first such studies, Shigematsu *et al*. [56] and Qu *et al*. [57] employed CNNs to RGB-D SOD. Although their network architectures are relatively straightforward, the performance is significantly improved. Subsequently, many works based on novel technologies, such as uncertainty learning [27], collaborative learning [28], joint learning [19], attention mechanism [20], [25], [58], graph neural network [35] and generative adversarial network [24], [59], were proposed.

The result fusion strategy [60], [61], single-stream strategy [33], [62] and two-stream cross-modal fusion strategy are three popular strategies, especially the last one. For typical examples, Li *et al*. [16] proposed a cross-modal depth-weighted combination block to further enhance RGB features with depth features. In [17], they proposed a cross-modal cross-scale module to enhance RGB features across scale. Chen *et al*. [38], [63]–[67] exploited the cross-modal fusion strategy in all its aspects, and proposed some effective methods, including cross-view transfer and multiview fusion [63], complementarity-aware fusion module [64], multi-scale multi-path and cross-modal interactions [65], three-stream attention-aware network [66], cross-level feedback fusion [67], and disentangled cross-modal fusion [38]. Piao *et al*. [68] proposed the depth-induced multi-scale weighting module to fuse multi-level refined features. In [18], Piao *et al*. proposed an efficient network with an adaptive and attentive depth distiller. Zhou *et al*. [23] fused both five-level features from RGB and depth stream to directly capture the complementary information. Fan *et al*. [29] employed the bifurcated backbone strategy to a cascaded refinement network, resulting in an effective teacher and student network. Chen *et al*. [31] fused depth features and RGB features alternately and progressively in an asymmetric network. Li *et al*. [32] designed the cross-modality feature modulation module to learn pixel-wise affine transformation parameters from the depth features for modulating RGB features. In [30], the dynamic dilated pyramid module was proposed to generate region-aware dynamic filters to decode RGB features. Zhang *et al*. [34] used a flow ladder module and a lightweight depth network for accurate saliency detection. Zhang *et al*. [69] introduced holistic aggregation paths and an extra bottom-up decoder network to enrich the cross-modal feature aggregation. Zhao *et al*. [70] proposed an effective gated dual branch structure to control the transmission of the valuable context information from the encoder to the decoder. Based on Siamese architecture, Fu *et al*. [71] combined the joint learning with the densely cooperative fusion to effectively exploit cross-modal complementarity. In [72], Chen *et al*. proposed the first 3D CNNs-based method for RGB-D SOD. Zhang *et al*. [73] extended the common foreground-first attention and proposed the bilateral attention mechanism, including foreground-first and background-first attention. The above deep learning-based methods have largely boosted the study of RGB-D SOD, however, the quality of depth map remained ignored.

Differently, Zhao *et al*. [50] trained a contrast-enhanced network to improve the quality of depth maps. This specific design made depth map much clearer and its regions more consistent, resulting in good performance. Similarly, Chen *et al*. [74] proposed an enhancement-and-fusion framework, which first generates a hint map to resolve the low-quality issue of depth maps. Chen *et al*. [21] introduced the two-phase depth estimation, which consists of coarse-level and second-round depth estimation, to large the depth differences of depth maps. The enhanced depth maps benefited subsequent selective saliency fusion. Fan *et al*. [9] designed the depth depurator unit to determine the quality of depth maps and discarded the poor ones in the pipeline. This method directly eliminates the distractors in poor depth maps at the source. Zhang *et al*. [26] selected useful regions from depth and RGB features in the cross-modal attention unit. The region selection operation was embedded into the network, and it was trained in an adaptive manner.

Inspired by the previous studies mentioned above, in this work, we focus on reducing the distractors in depth maps and propose a simple yet effective solution named HAINet, which is equipped with multiple novel HAIMs. Different from [21], [50], which separate the improvement of depth map quality and the inference of saliency map, our HAINet takes both into account and integrate them for RGB-D SOD. Instead of abandoning depth maps with poor quality as in [9], our HAINet aims to mine as much information as possible from depth maps, even from poor-quality ones. Compared with [26], which operates region selection in parallel, our HAINet performs RGB-depth-RGB interactions in a more effectively hierarchical and alternate manner. Thus, with all these advantages combined together, our HAINet can significantly overcome the negative effects of depth maps with poor quality and can successfully reason specific salient objects in the encoding-reasoning architecture.

C. Salient Object Detection on Multiple Visual Media

1) Image Salient Object Detection: Saliency detection was first developed on images by Itti *et al*. [75]. With the development of the last two decades, tremendous image SOD methods [6]–[8] have been proposed and widely used in other tasks. Traditional image SOD methods focus on utilizing low-level hand-crafted features, such as color, contrast, and object prior. Cheng *et al*. [76] proposed the famous global contrast based SOD method. Liu *et al*. [77] proposed a novel saliency tree framework for SOD. Wang *et al*. [78] introduced regional object-sensitive descriptors, including the objectness descriptor and the image-specific backgroundness descriptor, into SOD. Deep-learning based methods focus on encoding high-level semantic features and achieve superior

performance in image SOD. Wang *et al*. [79] proposed an iterative refinement strategy for image SOD and employed it in a recurrent fully convolutional network. Hou *et al*. [80] introduced the deep supervision and short connections, which have a great influence on subsequent methods, into image SOD. Huang *et al*. [81] developed a multi-level feature integration and multi-scale feature fusion based network for SOD. Liu *et al*. [82] proposed the pixel-wise contextual attention mechanism to select informative context regions for each pixel, which improves the pixel-level accuracy in the generated saliency map. In summary, RGB-D SOD methods are influenced by image SOD methods, and the strategies such as contrast and object prior are often applied in RGB-D SOD methods. However, the difference is that the key to RGB-D SOD lies in how to effectively mine the complementary information in the depth map.

2) Light Field Salient Object Detection: RGB-D image pair is the 3D data, while light field is the 4D data. The Light Field (LF) can record multiple viewpoints in one single shot, generating many 2D images with different viewpoint angles. These images imply with depth information and 3D shape information, and they can synthesize focal stack maps, depth maps and all-focus images.

Traditional methods for LF SOD are also based on hand-crafted features. Li *et al*. [83] proposed the first LF SOD method which exploits focusness, objectness and background prior. Later, Li *et al*. [84] developed a universal SOD framework based on weighted sparse coding to handle 2D, 3D and 4D data. Moreover, Zhang *et al*. [85] extracted the contrast saliency and background slice on depth map, all-focus map and focal stack maps. Zhang *et al*. [86] first generated the saliency priors in color, depth and flow cues from light-field images, then they enhanced saliency maps by location prior and refined them with the structure cue. These methods gradually excavated the depth and multi-view information of light field image and improved the accuracy of LF SOD.

Recently, the deep learning-based methods have significantly promoted the development of LF SOD. Wang *et al*. [87], Piao *et al*. [88], and Zhang *et al*. [89] successively proposed three light field saliency detection datasets, which contain 1,465, 1,580 and 1,462 images, respectively. Wang *et al*. [87] focused on effectively combining features of all-focus images and focal stacks in a two-stream network. Piao *et al*. [88] decomposed SOD into light field synthesis task and light-field-driven saliency detection task. Memory-oriented decoder is proposed by Zhang *et al*. [89] for deeply exploring and comprehensively exploiting internal correlation of focal slices. Notably, Zhang *et al*. [90] pointed out that these methods [87]–[89] were based on focal stacks and all-focal images, and they ignored the angular information in raw light field data. So, a method based on multiple viewpoint angles was proposed for LF SOD. Overall, light field images provide more focusness information and angular information than RGB-D images which only contain limited distance information. However, this also raises the problem of how to best integrate and utilize multi-modal information of light field images. We believe that the massive RGB-D SOD methods can provide some reference for LF SOD.

3) RGB-T Salient Object Detection: RGB-T salient object detection [91] is also a multi-modal SOD task, which is similar to the RGB-D SOD task. Specifically, 'T' refers to thermal infrared image, which is captured by thermal infrared camera. Compared with depth map, thermal infrared image is not limited by distance, but is easily affected by ambient temperature and has a lower resolution. Wang *et al*. [91] and Tang *et al*. [92] made the first attempt to RGB-T SOD, and proposed the first RGB-T SOD dataset, namely VT821 dataset. Tu *et al*. [93] performed multi-modal multi-scale manifold ranking on the superpixel-based graph, achieving saliency calculation. Recently, Tu *et al*. [94] created a new dataset, called VT1000, which contains 1,000 aligned RGB and thermal image pairs, and proposed a novel graph learning based method. Similar to most RGB-D SOD methods, Zhang *et al*. [95] proposed a multi-level feature fusion based network for RGB-T SOD. They designed a adjacent-depth features combination module for features extraction and a multi-branch group fusion module for capturing complementary features. Tu *et al*. [96] proposed an effective multi-interaction encoder-decoder network to explore the cross-modal complementarity among different modalities, different layers and local-global information. Considering the similarities between RGB-D SOD and RGB-T SOD tasks, we think that an excellent multi-modality architecture should perform well on both tasks.

III. Methodology

In this section, we elaborate the proposed Hierarchical Alternate Interaction Network (HAINet). In Sec. III-A, we present the overview and motivation of our HAINet. In Sec. III-B, we describe the details of Hierarchical Alternate Interaction Module (HAIM). In Sec. III-C, we provide the implementation details.

A. Network Overview and Motivation

The proposed HAINet is based on the encoding-reasoning architecture. As illustrated in Fig. 2, HAINet consists of feature encoding, cross-modal alternate interaction and saliency reasoning.

1) Feature Encoding: Considering the computation efficiency, we employ the relatively shallow VGG-16 [97] for feature encoding and remove the last max-pooling layer and three fully connected layers. As shown in Fig. 2, the RGB image and depth map are encoded separately through the two-stream modified VGG-16. These encoded blocks of RGB image and depth map are denoted by R-FE$^{(t)}$ and D-FE$^{(t)}$ ($t \in \{1, 2, 3, 4, 5\}$ is the block index), respectively. The effective feature encoding structure is widely used in recent RGB-D SOD methods [16], [17], [32], [63], [66]. Notably, for each block, we only operate on the feature map of the last convolutional layer, *i.e*., $\mathbf{F}_R^t$ of R-FE$^{(t)}$ and $\mathbf{F}_D^t$ of D-FE$^{(t)}$. The input resolutions of RGB image and depth map are set to $352 \times 352 \times 3$ and $352 \times 352 \times 1$, respectively.

2) Hierarchical Alternate Interaction Module: For the cross-modal RGB features and depth features, the interactions between them are crucial. Several common interactions in

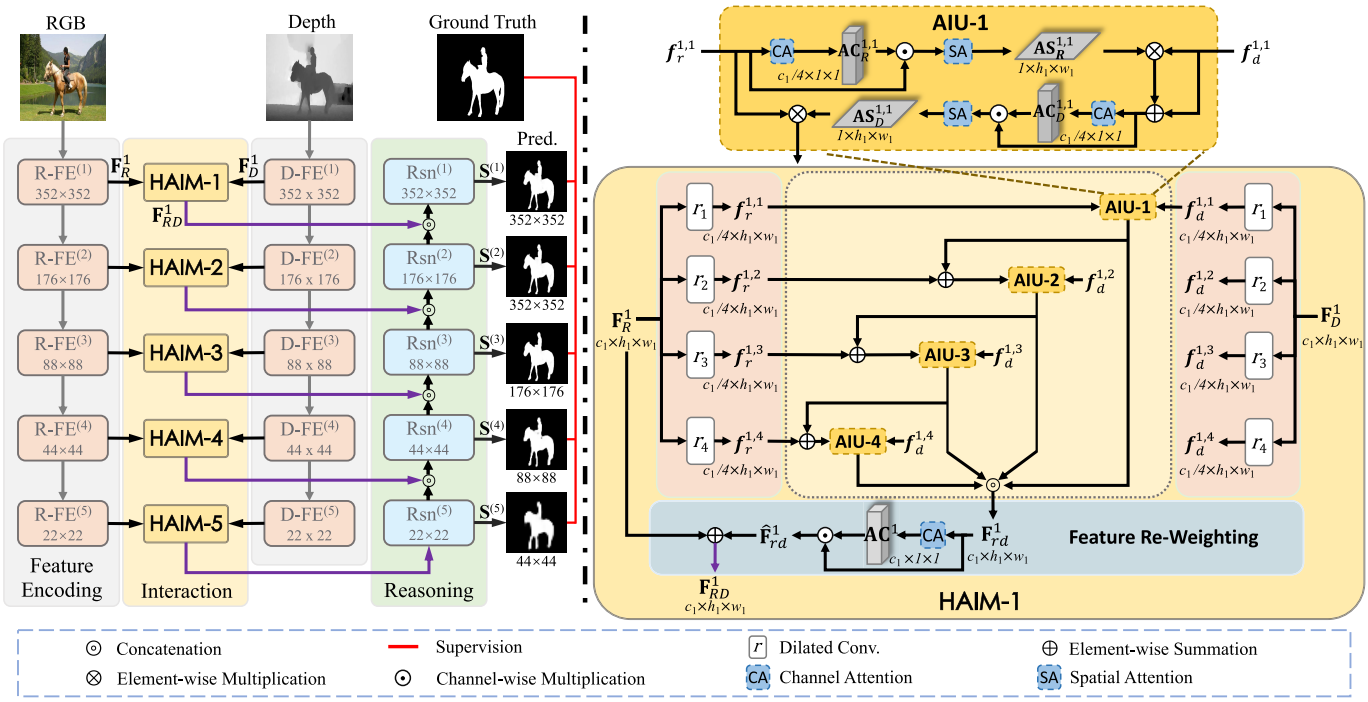


Fig. 2. Pipeline of the proposed HAINet. The left part shows that our HAINet contains three key stages: feature encoding, cross-modal alternate interaction and saliency reasoning. First, the feature encoding network extracts features from RGB image and depth map. Then, these cross-modal features are fed to Hierarchical Alternate Interaction Modules (HAIMs) for purifying depth features and enhancing RGB features (details of HAIM are shown on the right part). Finally, the output features of HAIMs are transferred to the saliency reasoning network for highlighting salient objects. Notably, at the training phase, the pixel-level supervision from the ground truth is attached to saliency reasoning network.

previous methods, such as multiplication-summation [16], [17] and concatenation-attention [66], [67], have been widely adopted. For multiplication-summation, depth features are used to modulate RGB features for feature enhancement; for concatenation-attention, the channel attention is applied to the concatenated features for feature screening. Different from previous methods, we aim at modeling an effective interaction which can filter distractors in depth features.

As shown in the bottom row of Fig. 1, although distance clues in the poor depth map may lead confusion, the RGB image shows color contrast. Thus motivated, we attempt to adopt RGB features for distractors filtering in depth features and propose the attention-multiplication interaction. Then, we employ the attention-multiplication interaction for distractors filtering, and we further enhance RGB features with the purified depth features similar to [16], [17]. Hence, we name the above interactions as modulation-feedback mechanism, and implement it in the Alternate Interaction Unit (AIU).

Moreover, we employ AIU in a hierarchical structure, which processes features with four parallel dilated convolutions [98]. Dilated rates of these dilated convolutions are incremental, which can capture local and global contexts. We believe that the kernel complementarity within a feature scale is necessary. As shown in Fig. 2, starting from the first branch, the output of former AIU will participate in the next modulation-feedback processing in the latter AIU, named progressive fusion. The hierarchical structure perfectly fits AIUs, generating features with rich contextual information.

To further effectively screen features and keep the most valuable information, we design an adaptive feature re-weighting operation to pay attention to four groups of enhanced RGB features discriminatively. With all these novel components working together, the proposed HAIM effectively overcomes the distractors in depth features and meanwhile fuses useful information from RGB and depth features. We introduce HAIM in detail in Sec. III-B, and examine the effectiveness of HAIM and its components in Sec. IV-C.

3) Saliency Reasoning: The saliency reasoning network is designed corresponding to the feature encoding network. These reasoning blocks are represented as $\mathrm{Rsn}^{(t)}$ ($t \in \{1, 2, 3, 4, 5\}$ is the block index). Each $\mathrm{Rsn}^{(t)}$ block consists of two ($\mathrm{Rsn}^{(1\sim2)}$) or three ($\mathrm{Rsn}^{(3\sim5)}$) convolutional layers and one deconvolutional layer. It receives features from the former block and the corresponding HAIM for progressive reasoning, where the deconvolutional layer is used to gradually restore the resolution.

4) Hybrid Loss: At the training phase, we employ a hybrid loss, composed of the Binary Cross-Entropy (BCE) loss and the Intersection-Over-Union (IOU) loss [99], to effectively train our HAINet. The pixel-level BCE loss and map-level IOU loss complement each other. We arrange the hybrid loss immediately after each $\mathrm{Rsn}^{(t)}$ block, as shown in Fig. 2. This deep supervision [100] not only makes our HAINet converge quickly, but also improves the accuracy of saliency reasoning. We represent the formulation and ablation study of hybrid Loss in Sec. III-C and Sec. IV-C, respectively.

B. Hierarchical Alternate Interaction Module

Hierarchical Alternate Interaction Module is the key component of HAINet. It bridges the feature encoding network and saliency reasoning network, and is responsible for distractors filtering and enhancement in the cross-modal features. The details of **HAIM-1** is illustrated in the right part of Fig. 2. There are three main components in HAIM: hierarchical branches, alternate interaction unit and feature re-weighting operation. In the following, we elaborate HAIM based on these three main components.

1) Hierarchical Branches: The sizes of RGB features $\mathbf{F}_R^t$ and depth features $\mathbf{F}_D^t$ are both defined as $\mathbb{R}^{c_t \times h_t \times w_t}$. We perform four hierarchical dilated convolutions on $\mathbf{F}_R^t$ and $\mathbf{F}_D^t$ with incremental dilated rates. These operations can be formulated as:

$$\boldsymbol{f}_r^{t,i} = \text{conv}(\mathbf{F}_R^t; \mathbf{W}^{t,i}, r_i), \quad i \in \{1,2,3,4\}, \tag{1}$$

$$\boldsymbol{f}_d^{t,i} = \text{conv}(\mathbf{F}_D^t; \mathbf{W}^{t,i}, r_i), \quad i \in \{1,2,3,4\}, \tag{2}$$

where $\{\boldsymbol{f}_r^{t,i}, \boldsymbol{f}_d^{t,i}\} \in \mathbb{R}^{c_t/4 \times h_t \times w_t}$ are the hierarchical features of RGB branch and depth branch, $\text{conv}(*; \mathbf{W}^{t,i}, r_i)$ is the dilated convolution with parameters $\mathbf{W}^{t,i}$ (*i.e.*, 3×3 kernel) and dilated rate $r_{i\in\{1,2,3,4\}} = \{1,3,5,7\}$, and i is the branch index. These dilated convolutions do not increase the computation, but large the receptive field. $\boldsymbol{f}_r^{t,i}$ and $\boldsymbol{f}_d^{t,i}$ effectively capture local and global contexts of $\mathbf{F}_R^t$ and $\mathbf{F}_D^t$, and they will benefit subsequent interactions in AIU.

2) Alternate Interaction Unit: As the **AIU-1** shown in Fig. 2, there are several channel attention (CA) [101] and spatial attention (SA) operations in AIU, which are defined as follows:

$$\text{CA}(\boldsymbol{f}) = \text{FC}_\sigma(\text{FC}_\phi(\text{GMP}_s(\boldsymbol{f}))), \tag{3}$$

$$\text{SA}(\boldsymbol{f}) = \text{conv}_\sigma(\text{GMP}_c(\boldsymbol{f})), \tag{4}$$

where $\text{FC}_\sigma(\cdot)$ is the fully connected layer with sigmoid activation function, $\text{FC}_\phi(\cdot)$ is the fully connected layer with ReLU activation function, $\text{GMP}_s(\cdot)$ is the spatial-wise global max pooling operation, $\text{GMP}_c(\cdot)$ is the channel-wise global max pooling operation, $\text{conv}_\sigma(\cdot)$ is the convolutional layer with sigmoid activation function, and $\boldsymbol{f}$ is the input feature.

Specifically, in AIU, we first apply the CA operation to the RGB branch, achieving CA map $\mathbf{AC}_R^{t,i} \in \mathbb{R}^{c_t/4 \times 1 \times 1}$, which is used to enhance features from RGB branch. Then, we apply the SA operation to the enhanced RGB features, achieving SA map $\mathbf{AS}_R^{t,i} \in \mathbb{R}^{1 \times h_t \times w_t}$. Different from $\mathbf{AC}_R^{t,i}$, $\mathbf{AS}_R^{t,i}$ is used to modulate features from depth branch for distractors filtering. We also add the modulated depth branch features to the original ones for contrast enhancement, achieving the RGB-depth modulation features $\boldsymbol{f}_{rd}^{t,i}$. This modulation process can be computed as:

$$\boldsymbol{f}_{rd}^{t,i} = \text{SA}(\text{CA}(\boldsymbol{f}_r^{t,i}) \odot \boldsymbol{f}_r^{t,i}) \otimes \boldsymbol{f}_d^{t,i} \oplus \boldsymbol{f}_d^{t,i}, \tag{5}$$

where $\odot$ is the channel-wise multiplication, $\otimes$ is the element-wise multiplication, and $\oplus$ is the element-wise summation.

Then, we apply the CA and SA operation to $\boldsymbol{f}_{rd}^{t,i}$, achieving $\mathbf{AC}_D^{t,i} \in \mathbb{R}^{(c_t/4) \times 1 \times 1}$ and $\mathbf{AS}_D^{t,i} \in \mathbb{R}^{1 \times h_t \times w_t}$ one after another.

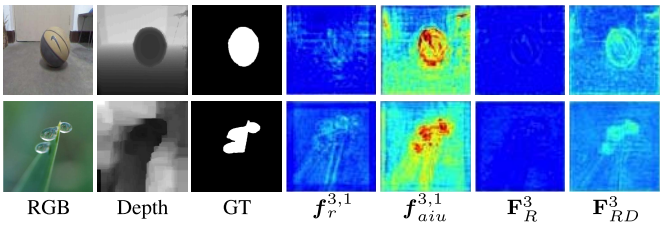


Fig. 3. Feature visualization in HAIM-3 and AIU-1 of HAIM-3. $f_r^{3,1}$ is the input RGB feature of AIU-1 in HAIM-3, while $f_{aiu}^{3,1}$ is the output feature of AIU-1 in HAIM-3. $\mathbf{F}_R^3$ is the input RGB feature of HAIM-3, while $\mathbf{F}_{RD}^3$ is the output feature of HAIM-3.

$\mathbf{AS}_D^{t,i}$ is in charge of enhancing RGB branch features, called depth-RGB feedback. The depth-RGB feedback features $\boldsymbol{f}_{dr}^{t,i}$ can be computed as:

$$\boldsymbol{f}_{dr}^{t,i} = \text{SA}(\text{CA}(\boldsymbol{f}_{rd}^{t,i}) \odot \boldsymbol{f}_{rd}^{t,i}) \otimes \boldsymbol{f}_r^{t,i}, \tag{6}$$

where $\boldsymbol{f}_{dr}^{t,i}$ is also the output feature of AIU-i, denoted as $\boldsymbol{f}_{aiu}^{t,i} \in \mathbb{R}^{c_t/4 \times h_t \times w_t}$.

In AIU, the interaction flow follows the RGB-depth-RGB manner, activating the modulation-feedback mechanism. Notably, as HAIM shown in Fig. 2, the previous $\boldsymbol{f}_{aiu}^{t,i}$ will participate in the next modulation-feedback processing. This progressive manner fuses the local and global features more effectively, and increases interactions among different branches. So, at the i-th branch, $\boldsymbol{f}_r^{t,i}$ in Eq. 5 and Eq. 6 should be re-written as follows:

$$\boldsymbol{f}_r^{t,i} = \begin{cases} \boldsymbol{f}_r^{t,i}, & i = 1 \\ \boldsymbol{f}_r^{t,i} \oplus \boldsymbol{f}_{aiu}^{t,i-1}, & i = 2,3,4. \end{cases} \tag{7}$$

After finishing all the modulation-feedback processing, we integrate features from four AIUs and obtain the contextual feature $\mathbf{F}_{rd}^t \in \mathbb{R}^{c_t \times h_t \times w_t}$ as follows:

$$\mathbf{F}_{rd}^t = \text{concat}\big(\boldsymbol{f}_{aiu}^{t,1}, \boldsymbol{f}_{aiu}^{t,2}, \boldsymbol{f}_{aiu}^{t,3}, \boldsymbol{f}_{aiu}^{t,4}\big), \tag{8}$$

where $\text{concat}(\cdot)$ is the cross-channel concatenation. With the above effective operations, $\mathbf{F}_{rd}^t$ contains rich contents of salient objects.

3) Feature Re-Weighting: Hierarchical convolutional layers with different receptive fields of HAIM have different responses to salient objects. Therefore, we decide to re-weight $\mathbf{F}_{rd}^t$ for further feature screening. For effectiveness, we modulate $\mathbf{F}_{rd}^t$ in a channel-wise manner with $\mathbf{AC}^t \in \mathbb{R}^{c_t \times 1 \times 1}$ generated from the adaptive CA operation, achieving more valuable feature $\hat{\mathbf{F}}_{rd}^t$. The channel-wise feature re-weighting operation can be formulated as:

$$\hat{\mathbf{F}}_{rd}^t = \text{CA}(\mathbf{F}_{rd}^t) \odot \mathbf{F}_{rd}^t. \tag{9}$$

$\hat{\mathbf{F}}_{rd}^t$ is purely reliant on the input feature $\mathbf{F}_{rd}^t$, which greatly enhances the flexibility of this adaptive manner.

Moreover, to preserve the original color contents, we stack $\mathbf{F}_R^t$ to $\hat{\mathbf{F}}_{rd}^t$ through the element-wise summation, achieving the output feature $\hat{\mathbf{F}}_{RD}^t \in \mathbb{R}^{c_t \times h_t \times w_t}$ of HAIM-t. Particularly, in HAIM-t, the size of output feature is half of the size of input feature, *i.e.*, $c_t \times h_t \times w_t$ *v.s.* $2 \times (c_t \times h_t \times w_t)$. HAIM

effectively reduces the size of input feature, and enhances the valuable contents of salient objects of output feature.

In Fig. 3, we visualize features[1] in HAIM-3 and AIU-1 of HAIM-3 to verify the effectiveness of AIU and HAIM. Concretely, in AIU-1 of HAIM-3, $\boldsymbol{f}_r^{3,1}$ is the input RGB feature and $\boldsymbol{f}_{aiu}^{3,1}$ is the output feature. The salient object in $\boldsymbol{f}_r^{3,1}$ is unrecognizable. After the RGB-depth-RGB interaction, the salient object is successfully highlighted in $\boldsymbol{f}_{aiu}^{3,1}$, which visually demonstrates that our modulation-feedback mechanism in AIU is effective. In HAIM-3, $\mathbf{F}_R^3$ is the input RGB feature of HAIM-3 and $\mathbf{F}_{RD}^3$ is the output feature of HAIM-3. The salient object in $\mathbf{F}_R^3$ almost disappears, especially in the bottom example of Fig. 3. With the assistance of three main components in HAIM, the salient regions are accurately highlighted in $\mathbf{F}_{RD}^3$, which proves that the combination of the three important components in our HAIM is reasonable and effective. In summary, our AIU and HAIM can overcome the low-quality situation of depth map, and produce high-quality features for subsequent salient object prediction.

4) Embedding HAIMs Into Encoding-Reasoning Architecture: Finally, we embed five HAIMs into the encoding-reasoning architecture, building the novel HAINet. Concretely, HAIM receives the cross-modal features from feature encoding network and produces valuable features for saliency reasoning network. It is worth noting that HAINet is extremely effective, running at 43 *fps* in our experiments, due to the seamless integration of HAIMs and network architecture.

C. Implementation Details

1) Total Loss Function: As shown in Fig. 2, each $\text{Rsn}^{(t)}$ block reasons a predicted saliency map, denoted as $\mathbf{S}^{(t)}$. Each $\mathbf{S}^{(t)}$ is supervised by the GT ($\mathbf{G}$) with the hybrid BCE loss and IOU loss for enhancing content representation at the training phase. We denote the total loss function as $\mathbb{L}$, which can be formulated as:

$$\mathbb{L} = \sum_{t=1}^{5} \left(\mathcal{L}_{bce}^{(t)}(up(\mathbf{S}^{(t)}), \mathbf{G}) + \mathcal{L}_{iou}^{(t)}(up(\mathbf{S}^{(t)}), \mathbf{G}) \right), \quad (10)$$

where $\mathcal{L}_{bce}^{(t)}(\cdot,\cdot)$ represents the BCE loss, $\mathcal{L}_{iou}^{(t)}(\cdot,\cdot)$ represents the IOU loss, and $up(\cdot)$ represents the bilinear upsampling that unsamples $\mathbf{S}^{(t)}$ to the same resolution as $\mathbf{G}$.

2) Network Training Protocol: We implement the proposed HAINet by PyTorch [102] with an NVIDIA Titan X GPU. Following [17], [63], the proposed HAINet is trained on a composite training set, including 1,400 samples from NJU2K [43] dataset and 650 samples from NLPR [44] dataset. At the training phase, all the training triplets first are resized to 352×352, and then they are augmented by randomly flipping, clipping and rotating. The initial parameters of the feature encoding network are adopted from the pre-trained VGG-16 model [97]. The normal distribution [103] is employed to initialize the parameters of all the newly added layers. The initial learning rate is set to 10^{-4}, which will be divided by 10 when training loss reaches a flat. We use the Adam optimizer [104] to train our HAINet with batch size of 5 for 59 epochs.

[1]We first compute the element-wise sum of feature maps, and then element-wisely divide by the number of feature maps, obtaining the average of feature maps for visualization.

IV. Experimental Results

A. Datasets and Evaluation Metrics

1) Datasets: We evaluate the proposed method on seven public benchmark datasets in this article.

STEREO [51] is the first stereoscopic image SOD dataset. It contains 1,000 pairs of binocular images, which are mainly collected from the Internet with coarse depth quality.

NJU2K [43] includes 2,003 stereo image pairs with various resolutions. Among these image pairs, 1,400 are used as training set, 100 are used as validation set, and the remaining are used as testing set.

DES [42] is a small dataset comprises 135 simple RGB-D images, with only one object in each image, from seven indoor scenes.

NLPR [44] consists of 1,000 images from 11 types of indoor and outdoor scenes. Among them, 650 images are used as training set, 50 images as validation set, and the remaining 300 images as testing set.

SIP [9] is the recently released dataset containing 929 images, which are designed for salient person detection and captured by a smart phone with high-quality depth map.

DUTLF-Depth [68] is a recent popular dataset for RGB-D SOD, and it consists of 1,200 images from 800 indoor and 400 outdoor scenes, which are challenging and complex.

ReDWeb-S [105] consists of 3,179 images from various real-world scenes, and it is a new large-scale challenging dataset for RGB-D SOD with high-quality depth maps.

2) Evaluation Metrics: S-measure ($\mathcal{S}_\lambda$, $\lambda = 0.5$) [106], maximum F-measure ($\mathcal{F}_\beta$, $\beta^2 = 0.3$) [107], maximum E-measure ($\mathcal{E}_\xi$) [108], weighted F-measure ($\mathcal{F}_\beta^w$, β^2 z 1) [109] and Mean Absolute Error (MAE, $\mathcal{M}$) are five widely used evaluation metrics in RGB-D SOD. The above metrics evaluate performance from multiple different aspects. We adopt them to quantitatively evaluate the performance of our method and other compared methods. We use the evaluation tool[2] provided by Fan *et al*. [9] to evaluate all methods.

S-measure focuses on measuring the structural similarity, which considers region-aware and object-aware structural similarity simultaneously. **F-measure** expresses different preferences for Precision (P) and Recall (R), which are defined as $\text{P} = \frac{|\mathbf{S}^{(1)} \cap \mathbf{G}|}{|\mathbf{S}|}$ and $\text{R} = \frac{|\mathbf{S}^{(1)} \cap \mathbf{G}|}{|\mathbf{G}|}$, respectively; where $\mathbf{S}^{(1)}$ and $\mathbf{G}$ denote respectively the predicted saliency map and ground truth. We pay more attention to precision in the article. **E-measure** is based on the cognitive characteristics of human vision system, which measures the local pixel-level errors and the global image-level errors together. **Weighted F-measure** measures the predicted pixels of saliency map according to their location and their neighborhood, which extends the basic F-measure to non-binary values. **MAE** measures the pixel-level errors between saliency map and ground truth.

[2]http://dpfan.net/d3netbenchmark/

TABLE I

QUANTITATIVE PERFORMANCE COMPARISON OF OUR METHOD AND OTHER 12 STATE-OF-THE-ART CNN-BASED METHODS, WHICH ARE TRAINED ON NJU2K [43] AND NLPR [44], ON FIVE DATASETS, INCLUDING STEREO [51], NJU2K [43], DES [42], NLPR [44], AND SIP [9], WITH S-MEASURE, MAXIMUM F-MEASURE, MAXIMUM E-MEASURE, WEIGHTED F-MEASURE AND MAE. WE ALSO REPORT THE FRAMES PER SECOND (FPS). ↑ AND ↓ INDICATE LARGER AND SMALLER IS BETTER, RESPECTIVELY. THE TOP THREE RESULTS ARE MARKED IN **RED**, **BLUE** AND **GREEN**

Models	FPS↑	STEREO [51]					NJU2K-T [43]					DES [42]					NLPR-T [44]					SIP [9]				
		$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$
DF$_{17}$ [57]	0.1	.757	.757	.847	.549	.141	.763	.804	.864	.591	.141	.752	.766	.870	.518	.093	.802	.778	.880	.584	.085	.653	.657	.759	.406	.185
CTMF$_{18}$ [63]	2	.848	.831	.912	.698	.086	.849	.845	.913	.720	.085	.863	.844	.932	.686	.055	.860	.825	.929	.679	.056	.716	.694	.829	.535	.139
PCF$_{18}$ [64]	17	.875	.860	.925	.778	.064	.877	.872	.924	.803	.059	.842	.804	.893	.714	.049	.874	.841	.925	.762	.044	.842	.838	.901	.768	.071
AFNet$_{19}$ [61]	33	.825	.823	.887	.752	.075	.772	.775	.853	.696	.100	.770	.728	.881	.641	.068	.799	.771	.879	.693	.058	.720	.712	.819	.617	.118
MMCI$_{19}$ [65]	20	.873	.863	.927	.760	.068	.858	.852	.915	.738	.079	.848	.822	.928	.650	.065	.856	.815	.913	.676	.059	.833	.818	.897	.712	.086
TANet$_{19}$ [66]	14	.871	.861	.923	.787	.060	.878	.874	.925	.804	.060	.858	.827	.910	.739	.046	.886	.863	.941	.780	.041	.835	.830	.895	.748	.075
CPFP$_{19}$ [50]	6	.879	.874	.925	.817	.051	.878	.877	.923	.828	.053	.872	.846	.923	.787	.038	.888	.867	.932	.813	.036	.850	.851	.903	.788	.064
D3Net$_{19}$ [9]	20	.891	.881	.930	.815	.054	.895	.889	.932	.833	.051	.904	.885	.946	.831	.030	.906	.885	.946	.826	.034	.864	.862	.903	.793	.063
***cm*SalGAN$_{20}$** [24]	1	.896	.888	.932	.828	.050	.903	.896	.940	.846	.046	.912	.898	.942	.839	.028	.922	.907	.957	.855	.027	.865	.864	.906	.795	.064
ICNet$_{20}$ [16]	13	.903	.898	.942	.844	.045	.894	.891	.926	.843	.052	.920	.913	.960	.867	.027	.923	.908	.952	.864	.028	.854	.857	.903	.791	.069
CMWNet$_{20}$ [17]	7	.905	.901	.944	.847	.043	.903	.902	.936	.857	.046	.934	.930	.969	.888	.022	.917	.903	.951	.856	.029	.867	.874	.913	.811	.062
UC-Net$_{20}$ [27]	17	.903	.899	.944	.867	.039	.897	.895	.936	.868	.043	.934	.930	.976	.908	.019	.920	.903	.956	.878	.025	.875	.879	.919	.836	.051
Ours	43	.907	.906	.944	.866	.040	.912	.915	.944	.883	.038	.935	.936	.973	.910	.018	.924	.915	.960	.887	.024	.880	.892	.922	.842	.053

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

1) Comparison Methods: We compare our method with 19 state-of-the-art CNN-based RGB-D SOD methods, including DF [57], CTMF [63], PCF [64], AFNet [61], MMCI [65], TANet [66], CPFP [50], D3Net [9], *cm*SalGAN [24], ICNet [16], CMWNet [17], UC-Net [27], DMRA [68], A2dele [18], FRDT [36], S^2MA [20], SSF [26], DANet [33], and CoNet [28]. Specifically, the training set of the first 12 methods (DF [57], CTMF [63], PCF [64], AFNet [61], MMCI [65], TANet [66], CPFP [50], D3Net [9], *cm*SalGAN [24], ICNet [16], CMWNet [17], and UC-Net [27]) is the same as that of our method, that is, a subset of NJU2K [43] and NLPR [44]. The training set of the last 7 methods (DMRA [68], A2dele [18], FRDT [36], S^2MA [20], SSF [26], DANet [33], and CoNet [28]) consists of a subset of NJU2K [43], NLPR [44] and DUTLF-Depth [68]. Following [33], [68], [105], we retrain our method with the training set of NJU2K [43], NLPR [44] and DUTLF-Depth [68]. For a fair comparison, the saliency maps of all compared methods are provided by authors[3] or obtained by running their released codes. Notably, the performance of DMRA [68], *cm*SalGAN [24], FRDT [36], SSF [26] and A2dele [18] in the original papers are tested on the STEREO-797 dataset which contains 797 images. We retest these methods on STEREO dataset with 1,000 images and report their performance in this article.

2) Quantitative Comparison: For the first 12 CNN-based methods, we report the quantitative performance comparison of our method and them on five datasets, including STEREO [51], NJU2K [43], DES [42], NLPR [44], and SIP [9], in terms of five metrics in Tab. I. Our method performs the best on NJU2K and NLPR datasets, and shows competitive performance on STEREO, DES and SIP datasets. More concretely, for $\mathcal{F}_\beta$, our method consistently outperforms all compared methods. On STEREO dataset, our method performs almost similar with the best method in terms of $\mathcal{E}_\xi$, *e.g.*, 0.866 (ours) *v.s.* 0.867 (UC-Net). Compared with the second best method on NJU2K dataset, the percentage gain of our method reaches 1.0% for $\mathcal{S}_\lambda$, 1.4% for $\mathcal{F}_\beta$, 1.7% for $\mathcal{F}_\beta^w$, and 11.6% for $\mathcal{M}$. On DES dataset, our method achieves the minimum pixel-level error 0.018 in $\mathcal{M}$. On NLPR dataset, our method is better than the suboptimal method by 0.9% in $\mathcal{F}_\beta^w$. On SIP dataset, our method improves the performance by 1.3% in $\mathcal{F}_\beta$.

For the last 7 CNN-based methods, we report the quantitative performance comparison of our method with them on seven datasets, including STEREO [51], NJU2K [43], DES [42], NLPR [44], SIP [9], DUTLF-Depth [68], and ReDWeb-S [105], in terms of five metrics in Tab. II. The test set of DUTLF-Depth [68] dataset contains 400 samples, and the test set of ReDWeb-S [105] dataset contains 1,000 samples. Our method achieves the best performance on four datasets, including NJU2K, NLPR, SIP and ReDWeb-S, and shows competitive performance on STEREO, DES and DUTLF-Depth datasets.

3) Visual Comparison: We represent visual comparisons with seven latest CNN-based RGB-D SOD methods in Fig. 4. There are several challenging and complicated scenes for RGB-D SOD: 1) poor depth map with distractors (1^{st} and 2^{nd} rows), 2) fair depth map (3^{rd} and 4^{th} rows), 3) good depth map with distractors (5^{th} and 6^{th} rows), and 4) good depth map (7^{th} and 8^{th} rows).

[3]Specifically, the saliency maps of DMRA [68], A2dele [18], S^2MA [20] and SSF [26] on ReDWeb-S [105] are provided by Liu *et al.* [105] on https://github.com/nnizhang/SMAC.

TABLE II

QUANTITATIVE PERFORMANCE COMPARISON OF OUR METHOD AND OTHER 7 STATE-OF-THE-ART CNN-BASED METHODS, WHICH ARE TRAINED ON NJU2K [43], NLPR [44], AND DUTLF-DEPTH [68], ON SEVEN DATASETS, INCLUDING STEREO [51], NJU2K [43], DES [42], NLPR [44], SIP [9], DUTLF-DEPTH [68], AND REDWEB-S [105], WITH S-MEASURE, MAXIMUM F-MEASURE, MAXIMUM E-MEASURE, WEIGHTED F-MEASURE AND MAE. WE ALSO REPORT THE FRAMES PER SECOND (FPS). ↑ AND ↓ INDICATE LARGER AND SMALLER IS BETTER, RESPECTIVELY. THE TOP THREE RESULTS ARE MARKED IN **RED**, **BLUE** AND **GREEN**

Models	FPS↑	STEREO [51]					NJU2K-T [43]					DES [42]					NLPR-T [44]				
		$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$
DMRA$_{19}$ [68]	10	.835	.847	.911	.779	.066	.886	.886	.927	.847	.051	.900	.888	.943	.843	.030	.899	.879	.947	.838	.031
A2dele$_{20}$ [18]	120	.879	.879	.928	.846	.045	.871	.873	.915	.841	.051	.886	.872	.921	.836	.028	.898	.882	.944	**.857**	**.029**
FRDT$_{20}$ [36]	21	**.902**	**.899**	**.943**	**.852**	**.042**	**.898**	**.899**	.933	**.855**	.048	.900	.886	.938	.838	.030	**.914**	**.900**	**.950**	**.857**	**.029**
S^2**MA**$_{20}$ [20]	9	.890	.882	**.932**	.825	.051	.894	.889	.930	.842	.053	**.941**	**.935**	**.973**	**.892**	**.021**	**.915**	**.902**	**.953**	.852	.030
SSF$_{20}$ [26]	20	.887	.882	.931	.843	.046	**.899**	**.896**	.934	**.863**	**.043**	.905	.885	.940	**.851**	**.025**	**.914**	.896	**.953**	**.865**	**.026**
DANet$_{20}$ [33]	32	.892	.882	.930	.830	.047	.897	.893	**.936**	.853	**.046**	.905	.895	**.958**	.848	.028	.909	.894	.949	.850	.031
CoNet$_{20}$ [28]	34	**.905**	**.901**	**.947**	**.866**	**.037**	.894	.893	**.937**	.849	.047	**.910**	**.896**	.945	.849	.027	.907	.887	.945	.842	.031
Ours	43	**.909**	**.909**	**.947**	**.871**	**.038**	**.909**	**.909**	**.941**	**.879**	**.038**	**.929**	**.924**	**.967**	**.898**	**.019**	**.921**	**.908**	**.954**	**.881**	**.025**

Models	SIP [9]					DUTLF-Depth [68]					ReDWeb-S [105]				
	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\lambda\uparrow$	$\mathcal{F}_\beta\uparrow$	$\mathcal{E}_\xi\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$
DMRA$_{19}$ [68]	.806	.821	.875	.740	.085	.889	.898	.933	.853	.048	.592	.579	.721	.456	.188
A2dele$_{20}$ [18]	.829	.834	.889	.779	.070	.887	.892	.929	.864	.043	**.641**	**.603**	**.672**	**.528**	**.160**
FRDT$_{20}$ [36]	.867	.871	.910	.811	.061	**.910**	.919	**.948**	.878	.039	-	-	-	-	-
S^2**MA**$_{20}$ [20]	.872	.877	**.919**	.819	.057	.903	.900	.937	.862	.044	**.711**	**.696**	**.781**	**.621**	**.139**
SSF$_{20}$ [26]	**.874**	**.880**	**.921**	**.828**	**.053**	**.916**	**.924**	**.951**	**.895**	**.034**	.595	.558	.710	.455	.189
DANet$_{20}$ [33]	**.878**	**.884**	**.921**	**.829**	**.054**	.890	.895	.931	.847	.047	-	-	-	-	-
CoNet$_{20}$ [28]	.858	.867	.913	.803	.063	**.919**	**.927**	**.956**	**.891**	**.033**	-	-	-	-	-
Ours	**.886**	**.903**	**.927**	**.854**	**.048**	**.910**	**.920**	.944	**.883**	**.038**	**.724**	**.723**	**.794**	**.654**	**.132**

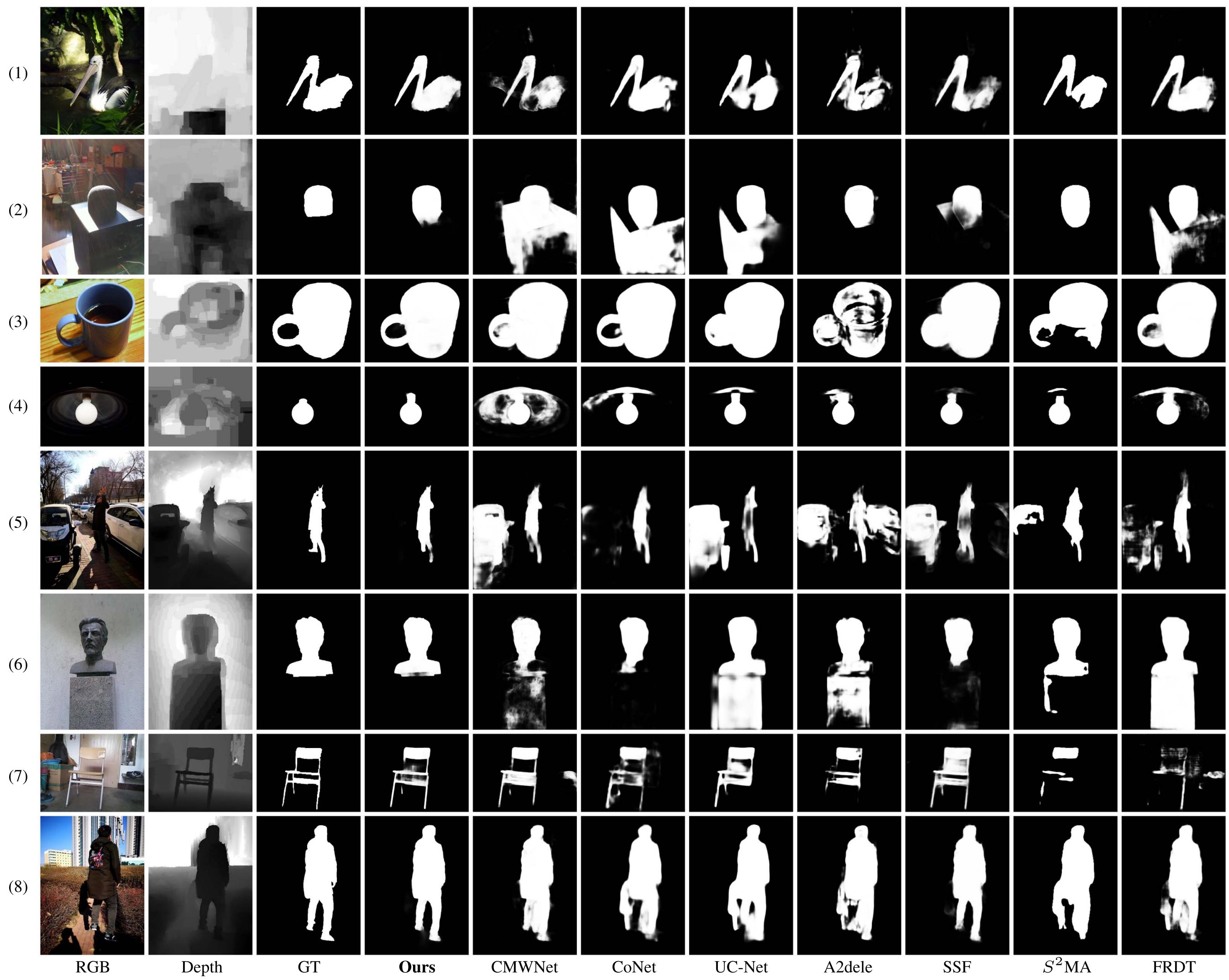


Fig. 4. Visual comparisons with seven latest CNN-based RGB-D SOD methods, including CMWNet [17], CoNet [28], UC-Net [27], A2dele [18], SSF [26], S^2MA [20], and FRDT [36].

For the first scene, the poor depth map fails to provide valuable distance cues, and even provides incorrect information (*i.e.*, distractors) of salient objects. Thanks to the power of the HAIM, the salient objects in our results of the first scene are relatively complete. For the second scene, our HAINet captures details of salient objects from the fair depth map.

TABLE III

ABLATION RESULTS ON CONFIRMING THE IMPORTANCE OF HAIM. SUM: REPLACING HAIM WITH ELEMENT-WISE SUMMATION AND CONVOLUTION OPERATIONS, AND CAT: REPLACING HAIM WITH CONCATENATION AND CONVOLUTION OPERATIONS. THE BEST RESULT IN EACH COLUMN IS **BOLD**

Models	STEREO [43]			NJU2K-T [43]			DES [42]		
	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$
Ours	**.907**	**.866**	**.040**	**.912**	**.883**	**.038**	**.935**	**.910**	**.018**
SUM	.894	.846	.045	.905	.871	.042	.915	.874	.024
CAT	.893	.846	.045	.902	.865	.043	.925	.891	.021

TABLE IV

ABLATION ANALYSES FOR THE RATIONALITY OF HIERARCHICAL STRUCTURE IN HAIM. WE PROVIDE FOUR HIERARCHICAL STRUCTURES WITH 1, 2, 4 (OURS) AND 8 BRANCHES. THE BEST RESULT IN EACH COLUMN IS **BOLD**

Branch	STEREO [43]			NJU2K-T [43]			DES [42]		
	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$
1	.897	.849	.046	.902	.865	.044	.914	.866	.026
2	.904	.861	.042	.904	.869	.043	.906	.865	.026
4 (Ours)	**.907**	**.866**	**.040**	**.912**	**.883**	**.038**	**.935**	**.910**	**.018**
8	.903	.861	.041	.909	.878	.039	.924	.893	.021

For example, the cup handle (4^{th} row) in our saliency map is finer. For the third scene, while the good depth map with distractors provides useful information about the salient objects, there is also some distractors involved, such as the car (5^{th} row) and the stone base (6^{th} row). Our HAINet effectively filters them and highlights salient objects. For the last scene, although the depth map has good quality, the RGB image is complicated. The saliency maps derived from our HAINet are still accurate.

4) Speed Comparison: We also report the Frame Per Second (FPS) in Tab. I and Tab. II. We borrow these speeds from their original papers and D3Net [9]. Our method reaches a real-time speed of 43 *fps*, ranking second among all 19 methods. Comparing with the fastest method A2dele, our method outperforms it by a large margin, *e.g.*, 2.6% to 12.0% in $\mathcal{F}_\beta$, as shown in Tab. II. Comparing with several recent state-of-the-art methods, such as UC-Net, CMWNet, S^2MA, SSF, and CoNet, our reasoning speed is much faster than them with comparable even better performance. Combining with quantitative results, visual presentations and speed comparison, we can find that our method is very competitive in RGB-D SOD community.

C. Ablation Studies

In this subsection, we provide comprehensive ablation studies on STEREO [51], NJU2K [43] and DES [42] to evaluate the contribution of each key component in our method. Specifically, we investigate 1) the importance of HAIM, 2) the rationality of hierarchical structure in HAIM, 3) the necessity of feature re-weighting in HAIM, 4) the effectiveness of alternate interaction in AIU, and 5) the usefulness of hybrid loss. We change one component each time and retrain variants with the same hyper-parameters and training set in Sec. III-C. Considering the alternate RGB-depth-RGB interaction is the key idea of our method, we name our complete method as *RDR* in Tab. V.

1) The Importance of HAIM: HAIM plays an important role in the proposed HAINet. To study its importance, we explore two relatively direct baseline variants: replacing HAIM with element-wise summation and convolution operations (*i.e.*, SUM) and replacing HAIM with concatenation and convolution operations (*i.e.*, CAT). As shown in Tab. III, the performance of both variants is severely damaged (*e.g.*, $\mathcal{F}_\beta^w$: 0.886 $\rightarrow$ 0.846 on STEREO of both variants, 0.883 $\rightarrow$ 0.865 on NJU2K of CAT, and 0.910 $\rightarrow$ 0.874 on DES of SUM). This confirms that our HAIM performs more expertly than direct operations (*i.e.*, SUM and CAT) for cross-modal interaction. With the assistance of HAIM, our HAINet overcomes the distractors of depth map with poor quality and captures the effective representation of salient objects.

Subsequently, we further recognize the contribution of each component, *i.e.*, hierarchical structure, alternate interaction unit and feature re-weighting operation, in HAIM.

2) The Rationality of Hierarchical Structure in HAIM: To validate the rationality of hierarchical structure in HAIM, we modify the number of branch and provide three variants: hierarchical structures with 1, 2 and 8 branches. The ablation results are reported in Tab. IV.

By comparing the results of 1, 2 and 4 (ours) branches, we discover that the performance basically increases with the addition of the number of branches among three datasets (*e.g.*, $\mathcal{M}$: 0.046 (1) $\rightarrow$ 0.042 (2) $\rightarrow$ 0.040 (4) on STEREO, 0.044 (1) $\rightarrow$ 0.043 (2) $\rightarrow$ 0.038 (4) on NJU2K, and 0.026 (1) $\rightarrow$ 0.026 (2) $\rightarrow$ 0.018 (4) on DES). This is because the more branches, the more local and global information interacts, resulting in better performance. However, this trend disappears in the variant with 8 branches (*e.g.*, $\mathcal{M}$: 0.040 (4) $\rightarrow$ 0.041 (8) on STEREO, 0.038 (4) $\rightarrow$ 0.039 (8) on NJU2K, and 0.018 (4) $\rightarrow$ 0.021 (8) on DES). The reason behind this is that the total channel number in each feature scale is fixed, and the channel number of each branch will decrease with the increase of branches, leading to the information loss of each branches (*e.g.*, $\mathbf{F}_R^1$ has 64 channels, if there are 8 branches in HAIM-1, $\boldsymbol{f}_r^{1,i}$ only has 8 channels). The adopted hierarchical structure with 4 branches achieves the balance between branch number and channel number, yielding the better results.

3) The Necessity of Feature Re-Weighting in HAIM: To explore the necessity of feature re-weighting in HAIM, we offer a variant which removes feature re-weighting in HAIM, *i.e.*, *w/o FRW*, and report the performance in Tab. V. From Tab. V, the performances are degraded (*e.g.*, $\mathcal{S}_\lambda$: 0.907 $\rightarrow$ 0.901 on STEREO, $\mathcal{F}_\beta^w$: 0.883 $\rightarrow$ 0.879 on NJU2K, and $\mathcal{M}$: 0.018 $\rightarrow$ 0.022 on DES), which confirm our adaptive feature re-weighting operation is necessary and the discriminate treatment for multiple $\boldsymbol{f}_{aiu}^{t,i}$ of $\mathbf{F}_{rd}^t$ is reasonable.

4) The Effectiveness of Alternate Interaction in AIU: In AIU, the alternate RGB-depth-RGB interaction is in charge

TABLE V

ABLATION ANALYSES ON CONFIRMING THE NECESSITY OF FEATURE RE-WEIGHTING IN HAIM, THE EFFECTIVENESS OF ALTERNATE INTERACTION IN AIU AND THE USEFULNESS OF HYBRID LOSS. *w/o FRW*: REMOVING FEATURE RE-WEIGHTING IN HAIM, *w/o D-R*: REMOVING THE LATTER DEPTH-RGB INTERACTION IN AIU, *w/o R-D*: REMOVING THE FORMER RGB-DEPTH INTERACTION IN AIU, *DRD*: REVERSING THE INPUT OF AIU (*i.e.*, RGB-DEPTH-RGB INTERACTION IS MODIFIED TO DEPTH-RGB-DEPTH INTERACTION), *w/o bce*: REMOVING BCE LOSS, *w/o iou*: REMOVING IOU LOSS, AND *w/o DS*: REMOVING DEEP SUPERVISION. THE BEST RESULT IN EACH COLUMN IS **BOLD**

Models	STEREO [43]			NJU2K-T [43]			DES [42]		
	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{M} \downarrow$
RDR (**Ours**)	**.907**	**.866**	**.040**	**.912**	**.883**	**.038**	**.935**	**.910**	**.018**
w/o FRW	.901	.855	.043	.910	.879	.039	.924	.892	.022
w/o D-R	.898	.854	.042	.906	.876	.042	.918	.879	.023
w/o R-D	.905	.864	**.040**	.906	.877	.041	.920	.887	.022
DRD	.899	.856	.043	.910	.878	.039	.928	.898	.021
w/o bce	.889	.859	.043	.899	.884	.041	.904	.869	.024
w/o iou	.906	.836	.047	.906	.853	.047	.926	.860	.027
w/o DS	.899	.853	.043	.906	.873	.042	.918	.877	.025

of distractors filtering (RGB-depth modulation interaction) and feature enhancement (depth-RGB feedback interaction). To investigate its effectiveness, we conduct three baselines: removing the latter depth-RGB interaction in AIU (*i.e.*, *w/o D-R*), removing the former RGB-depth interaction in AIU (*i.e.*, *w/o R-D*), and reversing the input of AIU (*i.e.*, RGB-depth-RGB interaction is modified to depth-RGB-depth interaction, called *DRD*). The structures of these three baselines are shown in Fig. 5.

As reported in Tab. V, the performance degradation of *w/o D-R* (*e.g.*, $\mathcal{S}_\lambda$: 0.907 $\rightarrow$ 0.898, $\mathcal{F}_\beta^w$: 0.866 $\rightarrow$ 0.854, and $\mathcal{M}$: 0.040 $\rightarrow$ 0.042 on STEREO) validates that the depth-RGB interaction benefits to highlight salient objects in RGB features. The performance of *w/o R-D* is also degraded (*e.g.*, $\mathcal{S}_\lambda$: 0.935 $\rightarrow$ 0.920, $\mathcal{F}_\beta^w$: 0.910 $\rightarrow$ 0.887, and $\mathcal{M}$: 0.018 $\rightarrow$ 0.022 on DES) which demonstrates the RGB-depth interaction contributes to distractors filtering in depth features. With the cooperation of RGB-depth and depth-RGB interaction (*i.e.*, the modulation-feedback mechanism), our method with the complete AIU achieves the best performance. In addition, the performance degradation of *DRD* (*e.g.*, $\mathcal{S}_\lambda$: 0.907 $\rightarrow$ 0.899, $\mathcal{F}_\beta^w$: 0.866 $\rightarrow$ 0.856, and $\mathcal{M}$: 0.040 $\rightarrow$ 0.043 on STEREO) demonstrates that our RGB-depth-RGB interaction is more effective than the depth-RGB-depth interaction for the cross-modal feature fusion.

Through the above detailed experiments and analyses, the superiority and effectiveness of the complete HAIM and its three key components are verified.

5) The Usefulness of Hybrid Loss: To investigate the usefulness of hybrid loss, we provide three variants: removing BCE loss (*i.e.*, *w/o bce*), removing IOU loss (*i.e.*, *w/o iou*), and removing deep supervision (*i.e.*, *w/o DS*). As shown in Tab. V, the results of *w/o bce* and *w/o iou* confirm that each loss can achieve acceptable performance alone, but more favorable performance can be attained when IOU loss and BCE loss are combined. We can also observe that *w/o iou* achieves the worst performance in $\mathcal{F}_\beta^w$ and $\mathcal{M}$ (*e.g.*, $\mathcal{F}_\beta^w$: 0.836 on STEREO and $\mathcal{M}$: 0.027 on DES), which confirms the IOU loss contributes significantly to pixel-level accuracy. The observation is valuable for RGB-D SOD and other related pixel-level prediction tasks. The deep supervision contributes to the performance, *e.g.*, it carries 0.017 performance improvement in $\mathcal{S}_\lambda$ on DES. Overall, our hybrid loss is pretty useful.

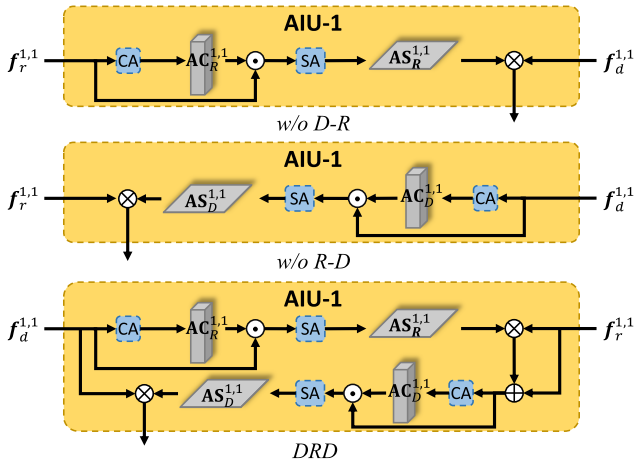


Fig. 5. Structures of three AIU variants. *w/o D-R*: removing the depth-RGB interaction in AIU, *w/o R-D*: removing the RGB-depth interaction in AIU, and *DRD*: reversing the input of AIU (*i.e.*, RGB-depth-RGB interaction is modified to depth-RGB-depth interaction).

D. Application to RGB-T SOD Task

In this subsection, we apply our method to a similar multi-modal SOD task, *i.e.*, RGB-T SOD, to demonstrate the effectiveness of our method. Specifically, following a recent work [96] for RGB-T SOD, we adopt VT1000 [94] dataset to retrain our method and test our method on VT821 [92] dataset. VT1000 [94] contains 1,000 samples of aligned RGB and thermal images, while VT821 [92] contains 821 samples.

We compare our method with five state-of-the-art RGB-T SOD methods, including MTMR [91], M3S-NIR [93], SGDL [94], ADF [110] and MIED [96], in terms of S-measure, F-measure, E-measure, weighted F-measure and MAE. Saliency maps of all compared RGB-T SOD methods are provided by authors.[4] The quantitative performance is reported in Tab. VI. Our method performs the best on VT821 dataset, which demonstrates the robustness and generalizability of our method for multi-modal SOD tasks.

E. Failure Cases and Analyses

In this subsection, we illustrate some failure cases of our HAINet and present analyses. As shown in Fig. 6, there are two challenging cases for our HAINet: 1) the salient objects with occlusion (the top two rows) and 2) the scenes with multiple salient objects (the bottom two rows). In the top two

[4]https://github.com/lz118/RGBT-Salient-Object-Detection

TABLE VI

QUANTITATIVE PERFORMANCE COMPARISON OF OUR METHOD AND FIVE STATE-OF-THE-ART RGB-T SOD METHODS ON VT821 [92] DATASET. THE BEST RESULT IN EACH ROW IS **BOLD**

Metric	MTMR_{18} [91]	M3S-NIR_{19} [93]	SGDL_{20} [94]	ADF_{20} [110]	MIED_{20} [96]	**Ours**
$\mathcal{S}_\lambda \uparrow$	0.725	0.723	0.765	0.810	0.849	**0.856**
$\mathcal{F}_\beta \uparrow$	0.690	0.738	0.735	0.752	0.809	**0.819**
$\mathcal{E}_\xi \uparrow$	0.812	0.837	0.839	0.839	0.887	**0.897**
$\mathcal{F}_\beta^w \uparrow$	0.462	0.407	0.583	0.626	0.775	**0.776**
$\mathcal{M} \downarrow$	0.108	0.140	0.085	0.077	0.046	**0.044**

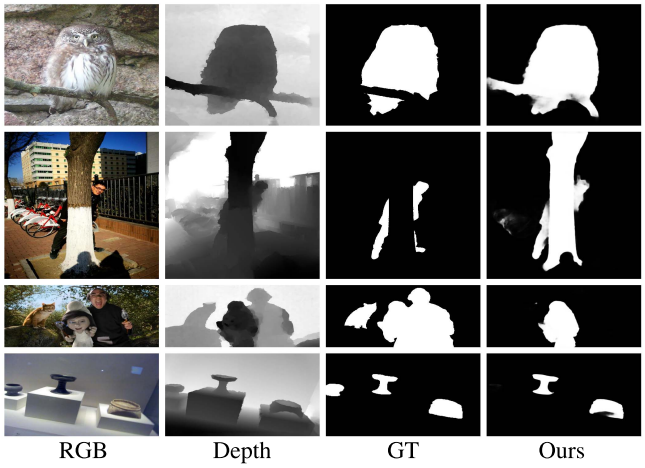


Fig. 6. Some failure cases of our HAINet. The top two rows show the salient objects with occlusion, and the bottom two rows show the scenes with multiple salient objects.

rows, the depth maps have good quality, but the salient objects are occluded. Our method fails to accurately highlight the salient objects and mistakenly highlight the occluded objects. The occluded objects are different from distractors in the depth maps and they are adjacent to salient objects, causing our method to mistake them for being part of salient objects. In the bottom two rows, there are multiple similar objects in RGB images and depth maps, but our method highlights only part of them. This is because different salient objects have different distance information in the depth map, which causes our method to mistake salient objects with far distance as distractions and thus suppress them.

V. CONCLUSION

In this article, we have proposed a novel and effective Hierarchical Alternate Interaction Network (HAINet) for RGB-D SOD. HAINet is based on the encoding-reasoning architecture, and is equipped with dedicated efficient Hierarchical Alternate Interaction Modules (HAIMs). In particular, HAIM is a vital medium for encoding and reasoning network, and is in charge of enhancing interactions between cross-modal features, *i.e.*, filtering distractors in depth features and enhancing the content representation of salient objects on RGB features. Thanks to the simplicity of modules, the effectiveness of the architecture and the tacit cooperation between them, HAINet achieves a fast reasoning speed of 43 *fps* on a single GPU. Comprehensive experiments, including comparison analyses and ablation studies, demonstrate that HAINet is competitive to 19 state-of-the-art RGB-D SOD methods and strikes a balance between precision and speed.

REFERENCES

[1] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. ICML*, Jul. 2015, pp. 597–606.
[2] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1473–1482.
[3] Y. Liu, Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu, and M.-M. Cheng, "Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 10, 2020, doi: 10.1109/TPAMI.2020.3023152.
[4] 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.
[5] G. Li *et al.*, "Personal fixations-based object segmentation with object localization and boundary preservation," *IEEE Trans. Image Process.*, vol. 30, pp. 1461–1475, Jan. 2021.
[6] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.
[7] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, vol. 5, no. 2, pp. 117–150, Jun. 2019.
[8] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 12, 2021, doi: 10.1109/TPAMI.2021.3051099.
[9] 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 Trans. Neural Netw. Learn. Syst.*, early access, Jun. 3, 2020, doi: 10.1109/TNNLS.2020.2996406.
[10] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "RGB-D salient object detection: A survey," *Comput. Vis. Media*, pp. 1–33, Jan. 2021, doi: 10.1007/s41095-020-0199-z.
[11] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, Jan. 2021.
[12] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.
[13] 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.
[14] X. Wang, H. Ma, X. Chen, and S. You, "Edge preserving and multi-scale contextual neural network for salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 121–134, Jan. 2018.
[15] X. Xiao, Y. Zhou, and Y.-J. Gong, "RGB-'D' saliency detection with pseudo depth," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2126–2139, May 2019.
[16] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, Mar. 2020.
[17] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 665–681.
[18] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9057–9066.
[19] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3049–3059.
[20] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13753–13762.

[21] 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 Trans. Image Process.*, vol. 29, pp. 4296–4307, Jan. 2020.

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

[23] X. Zhou, G. Li, C. Gong, Z. Liu, and J. Zhang, "Attention-guided RGBD saliency detection using appearance information," *Image Vis. Comput.*, vol. 95, Mar. 2020, Art. no. 103888.

[24] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo, "CmSalGAN: RGB-D salient object detection with cross-view generative adversarial networks," *IEEE Trans. Multimedia*, early access, May 28, 2020, doi: 10.1109/TMM.2020.2997184.

[25] D. Liu, K. Zhang, and Z. Chen, "Attentive cross-modal fusion network for RGB-D saliency detection," *IEEE Trans. Multimedia*, vol. 23, pp. 967–981, 2021, doi: 10.1109/TMM.2020.2991523.

[26] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3469–3478.

[27] J. Zhang *et al.*, "UC-NET: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8579–8588.

[28] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D salient object detection via collaborative learning," in *Proc. ECCV*, Aug. 2020, pp. 52–69.

[29] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. ECCV*, Aug. 2020, pp. 275–292.

[30] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 235–252.

[31] S. Chen and Y. Fu, "Progressively guided alternate refinement network for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 520–538.

[32] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "RGB-D salient object detection with cross-modality modulation and selection," in *Proc. ECCV*, Aug. 2020, pp. 225–241.

[33] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 646–662.

[34] M. Zhang, S. Xiao-Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, "Asymmetric two-stream architecture for accurate RGB-D saliency detection," in *Proc. ECCV*, Aug. 2020, pp. 646–662.

[35] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu, "Cascade graph neural networks for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 346–364.

[36] 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 *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4107–4115.

[37] J. Zhao, Y. Zhao, J. Li, and X. Chen, "Is depth really necessary for salient object detection?" in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1745–1754.

[38] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "RGBD salient object detection via disentangled cross-modal fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 8407–8416, Aug. 2020.

[39] G. Liao, W. Gao, Q. Jiang, R. Wang, and G. Li, "MMNet: Multi-stage and multi-scale fusion network for RGB-D salient object detection," in *Proc. ACM MM*, Oct. 2020, pp. 2436–2444.

[40] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection" *IEEE Trans. Image Process.*, early access, Nov. 3, 2020, doi: 10.1109/TIP.2020.3028289.

[41] W. Zhou, Y. Chen, C. Liu, and L. Yu, "GFNet: Gate fusion network with Res2Net for detecting salient objects in RGB-D images," *IEEE Signal Process. Lett.*, vol. 27, pp. 800–804, May 2020.

[42] 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)*, 2014, pp. 23–27.

[43] 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.

[44] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.

[45] 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.

[46] 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.

[47] H. Song, Z. Liu, H. Du, G. Sun, and C. Bai, "Saliency detection for RGBD images," in *Proc. 7th Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2015, pp. 1–4.

[48] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] A. Wang and M. Wang, "RGB-D salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 663–667, May 2017.

[56] 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.

[57] 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.

[58] C. Li *et al.*, "ASIF-net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 2021.

[59] Z. Liu, W. Zhang, and P. Zhao, "A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection," *Neurocomputing*, vol. 387, pp. 210–220, Apr. 2020.

[60] 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.

[61] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, May 2019.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.

[67] H. Chen, Y. Li, and D. Su, "Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4808–4820, Nov. 2020.

[68] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7253–7262.

[69] Y.-F. Zhang *et al.*, "Rethinking feature aggregation for deep RGB-D salient object detection," *Neurocomputing*, vol. 423, pp. 463–473, Jan. 2021.

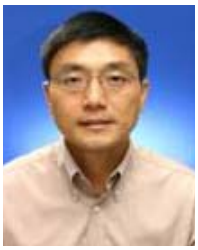
[70] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 35–51.
[71] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," 2020, *arXiv:2008.12134*. [Online]. Available: http://arxiv.org/abs/2008.12134
[72] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "RGB-D salient object detection via 3D convolutional neural networks," in *Proc. AAAI*, 2021, pp. 1–9.
[73] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1949–1961, Jan. 2021.
[74] Q. Chen *et al.*, "EF-net: A novel enhancement and fusion network for RGB-D saliency detection," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107740.
[75] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[76] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.
[77] 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.
[78] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 251–268, Jun. 2017.
[79] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, Jul. 2019.
[80] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.
[81] M. Huang, Z. Liu, L. Ye, X. Zhou, and Y. Wang, "Saliency detection via multi-level integration and multi-scale fusion neural networks," *Neurocomputing*, vol. 364, pp. 310–321, Oct. 2019.
[82] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Pixel-wise contextual attention learning for accurate saliency detection," *IEEE Trans. Image Process.*, vol. 29, pp. 6438–6451, Apr. 2020.
[83] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1605–1616, Aug. 2017.
[84] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5216–5223.
[85] J. Zhang, M. Wang, J. Gao, Y. Wang, X. Zhang, and X. Wu, "Saliency detection with a deeper investigation of light field," in *Proc. IJCAI*, 2015, pp. 2212–2218.
[86] J. Zhang, M. Wang, L. Lin, X. Yang, J. Gao, and Y. Rui, "Saliency detection on light field: A multi-cue approach," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 3, pp. 1–22, 2017.
[87] T. Wang, Y. Piao, H. Lu, X. Li, and L. Zhang, "Deep learning for light field saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8837–8847.
[88] Y. Piao, Z. Rong, M. Zhang, X. Li, and H. Lu, "Deep Light-field-driven saliency detection from a single view," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 904–911.
[89] M. Zhang, J. Li, W. Ji, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," in *Proc. NeurIPS*, Dec. 2019, pp. 898–908.
[90] Q. Zhang, S. Wang, X. Wang, Z. Sun, S. Kwong, and J. Jiang, "A multi-task collaborative network for light field salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jul. 30, 2020, doi: 10.1109/TCSVT.2020.3013119.
[91] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *Proc. Chin. Conf. Image Graph. Technol.*, Aug. 2018, pp. 359–369.
[92] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "RGBT salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4421–4433, Dec. 2020.
[93] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang, "M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Mar. 2019, pp. 141–146.
[94] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 160–173, Jan. 2020.
[95] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.
[96] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive siamese decoder for RGBT salient object detection," 2020, *arXiv:2005.02315*. [Online]. Available: http://arxiv.org/abs/2005.02315
[97] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, May 2015, pp. 1–14.
[98] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, May 2016, pp. 1–13.
[99] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.
[100] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.
[101] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[102] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, Dec. 2019, pp. 8024–8035.
[103] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[104] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, May 2015, pp. 1–15.
[105] N. Liu, N. Zhang, L. Shao, and J. Han, "Learning selective mutual attention and contrast for RGB-D saliency detection," 2020, *arXiv:2010.05537*. [Online]. Available: http://arxiv.org/abs/2010.05537
[106] 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.
[107] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.
[108] 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.
[109] 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.
[110] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "RGBT salient object detection: A large-scale dataset and benchmark," 2020, *arXiv:2007.03262*. [Online]. Available: http://arxiv.org/abs/2007.03262

Gongyang Li (Member, IEEE) 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 by EU FP7 Marie Curie Actions. He has published more than 200 refereed technical articles in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He was a TPC member/session chair in 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* and served as a Guest Editor for the special issue on Recent Advances in Saliency Models, Applications and Evaluations in *Signal Processing: Image Communication*.

Minyu Chen received the B.E. degree from the Zhejiang University of Technology, Zhejiang, China, in 2018. He is currently pursuing the M.E. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests include deep learning and image inpainting.

Zhen Bai received the B.E. degree from the Wuhan Huaxia University of Technology, Wuhan, China, in 2016, and the M.S. degree from the Zhengzhou University of Light Industry, Zhengzhou, China, in 2019. She is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. Her research interests include machine learning and saliency detection.



Weisi Lin (Fellow, IEEE) received the Ph.D. degree from King's College London, U.K. He served as the Laboratory Head for visual processing with the Institute for Infocomm Research, Singapore. He is currently a Professor with the School of Computer Engineering. His research interests include image processing, perceptual signal modeling, video compression, and multimedia communication. He has published more than 200 journal articles, more than 230 conference articles, filed 11 patents, and authored two books in the above areas.

He is a fellow of IET and an honorary fellow of the Singapore Institute of Engineering Technologists. He has been the Technical Program Chair of the IEEE ICME 2013, the PCM 2012, the QoMEX 2014, and the IEEE VCIP 2017. He has been an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE SIGNAL PROCESSING LETTERS, and the *Journal of Visual Communication and Image Representation*. He has been an invited/panelist/keynote/tutorial speaker for more than 20 international conferences. He was a Distinguished Lecturer of the Asia-Pacific Signal and Information Processing Association (APSIPA) from 2012 to 2013 and the IEEE Circuits and Systems Society from 2016 to 2017.

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 at College Park, College Park, in 2006. From 2000 to 2001, he was an Assistant Researcher with Microsoft Research Asia. From 2006 to 2007, he was a Postdoctoral Scientist with the University of California at Los Angeles. In 2007, he joined as a Research Scientist with Siemens Corporate Research; then, from 2008 to 2019, he was a Faculty Member of the Department of Computer Sciences, Temple University. In fall 2019, he joined as a SUNY Empire Innovation Professor with the Department of Computer Science, Stony Brook University. 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, NSF CAREER Award in 2014, Yahoo Faculty Research and Engagement Program Award in 2019, and Amazon AWS Machine Learning Research Award in 2019. He has served as an area chairs various times for CVPR and ECCV. He serves as Associate Editors for several journals including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI), *Pattern Recognition* (PR), and *Computer Vision and Image Understanding* (CVIU).