Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection

Gao, Wei; Liao, Guibiao; Ma, Siwei; Li, Ge; Liang, Yongsheng; Lin, Weisi

2021

Gao, W., Liao, G., Ma, S., Li, G., Liang, Y. & Lin, W. (2021). Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. IEEE Transactions On Circuits and Systems for Video Technology.

<https://dx.doi.org/10.1109/TCSVT.2021.3082939>

<https://hdl.handle.net/10356/150772>

<https://doi.org/10.1109/TCSVT.2021.3082939>

© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at:
<https://doi.org/10.1109/TCSVT.2021.3082939>.

Downloaded on 26 Apr 2023 13:25:36 SGT

Unified Information Fusion Network for Multi-Modal RGB-D and RGB-T Salient Object Detection

Wei Gao, *Member, IEEE*, Guibiao Liao, Siwei Ma, Ge Li, Yongsheng Liang, and Weisi Lin, *Fellow, IEEE*

Abstract—The use of complementary information, namely depth or thermal information, has shown its benefits to salient object detection (SOD) during recent years. However, the RGB-D or RGB-T SOD problems are currently only solved independently, and most of them directly extract and fuse raw features from backbones. Such methods can be easily restricted by low-quality modality data and redundant cross-modal features. In this work, a unified end-to-end framework is designed to simultaneously analyze RGB-D and RGB-T SOD tasks. Specifically, to effectively tackle multi-modal features, we propose a novel multi-stage and multi-scale fusion network (MMNet), which consists of a cross-modal multi-stage fusion module (CMFM) and a bi-directional multi-scale decoder (BMD). Similar to the visual color stage doctrine in the human visual system (HVS), the proposed CMFM aims to explore important feature representations in feature response stage, and integrate them into cross-modal features in adversarial combination stage. Moreover, the proposed BMD learns the combination of multi-level cross-modal fused features to capture both local and global information of salient objects, and can further boost the multi-modal SOD performance. The proposed unified cross-modality feature analysis framework based on two-stage and multi-scale information fusion can be used for diverse multi-modal SOD tasks. Comprehensive experiments ($\sim 92K$ image-pairs) demonstrate that the proposed method consistently outperforms the other 21 state-of-the-art methods on nine benchmark datasets. This validates that our proposed method can work well on diverse multi-modal SOD tasks with good generalization and robustness, and provides a good multi-modal SOD benchmark.

Index Terms—Dynamic cross-modal guided mechanism, RGB-D/RGB-T multi-modal data, information fusion, salient object detection.

I. INTRODUCTION

SALIENT object detection (SOD) aims to identify the most attractive regions in natural images. It can benefit

This work was supported by Ministry of Science and Technology of China - Science and Technology Innovations 2030 (2019AAA0103501), Natural Science Foundation of China (61801303 and 62031013), Guangdong Basic and Applied Basic Research Foundation (2019A1515012031), and Shenzhen Science and Technology Plan Basic Research Project (JCYJ20190808161805519). (*Corresponding author: Wei Gao, Guibiao Liao.*)

W. Gao, G. Liao and G. Li are with School of Electronic and Computer Engineering, Peking University, and also Peng Cheng Laboratory, Shenzhen, China (Email: gaowei262@pku.edu.cn, leno_biao@163.com, geli@pku.edu.cn).

S. Ma is with the Institute of Digital Media, Peking University, Beijing, China (Email: swma@pku.edu.cn).

Y. Liang is with the School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, China (Email: liangys@hit.edu.cn).

W. Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (Email: wslin@ntu.edu.sg).

A preliminary version of this work has appeared in ACM MM 2020 [1].

Fig. 1. Performance comparisons for different RGB-D and RGB-T SOD methods. Our method can generate better saliency results when it comes to the unreliable complementary modality data (rows 1, 4), challenging scenarios (rows 2, 5), and multiple objects with various sizes (rows 3, 6).

many image and video processing tasks, such as segmentation [2], [3], compression [4], visual tracking [5], and quality assessment [6], etc. Although SOD has made great progresses during recent years [7]–[15], they heavily rely on information from individual RGB images. It is still difficult to achieve accurate SOD results in some challenging and complex scenarios, e.g., the low color contrasts between salient objects and background, cluttered background interference, etc. With the widespread use of depth cameras (e.g., Kinect, iPhone XR, and Huawei Mate10) and infrared imaging devices, both the captured depth and thermal information have shown benefits to SOD performance improvements. Nevertheless, how to effectively implement cross-modality information fusion is still challenging, which can significantly affect the achievement of robust performance.

Depth information can provide additional geometric cues, such as spatial structure and 3D layout, and hence RGB-D SOD is becoming an attractive topic [16]–[23]. Although many previous works have attempted to explore the use of depth information for RGB-D SOD, some common limitations remain. (1) **Depth map quality issues:** Unfortunately, high-

quality depth maps are still scarce [18]. The inappropriate utilization of low-quality depth maps will deteriorate the performance of SOD. The recent work of D3Net [24] used depth depurator unit (DDU) to automatically filter out low-quality depth maps. However, it may encounter some challenging situations without depth information and only with RGB modality data as shown in the first row of Fig. 1. **(2) Cross-modal fusion effects:** As evidenced in Fig. 1, the ability to extract representative cross-modal features through fusion strategies, such as concatenation and simple convolutional layers, may not take full advantage of complementarity between modalities. **(3) Performance instability issues for differently-scaled objects:** As exemplified in the third row of Fig. 1, there are multiple salient objects with different sizes, and thus robustly locating these salient objects from RGB-D data is an important task that may be easily neglected.

In addition, integrating thermal infrared and RGB data has also shown its effectiveness in salient object detection [25]–[29]. Specifically, thermal infrared data can capture the radiated heat of objects and provide supplementary information for better detection when it comes to cluttered backgrounds or low illumination. Similar to RGB-D, RGB-T SOD is also a cross-modal fusion issue. It is important for SOD to capture the interrelationship and make use of the complementarity between the two different modalities. However, efficient fusion methods are still underexplored. Without sufficient consideration on the cross-modal data inconsistency and the robustness of multi-scale detection, directly fusing cross-modal features may generate suboptimal saliency results by the existing methods as shown in Fig. 1.

From the above analysis for the existing RGB-D and RGB-T SOD works, we can see that there are three main issues to be addressed in multi-modal SOD: (1) How to select and utilize available single modal features; (2) How to combine cross-modal information more effectively and accurately; (3) How to process the situations where multiple objects have very different sizes in a scene.

Due to the critical role of the cross-modality fusion in SOD, an effective fusion strategy is demanded. The renowned visual color stage doctrine [30]–[36] in the human visual system (HVS) explains the efficient selecting and adversarial combination processes for visual color perception. In fact, different modality data and information involved in SOD act in a similar manner [37]–[43]. Therefore, to effectively address the above-mentioned problems, inspired by the visual color stage doctrine, this paper proposes a novel multi-stage and multi-scale fusion network (MMNet) for better multi-modal RGB-D and RGB-T SOD performance. As shown in Fig. 2, it mainly consists of a cross-modal multi-stage fusion module (CMFM) and a bi-directional multi-scale decoder (BMD).

Unlike previous works [18], [24], [29] that employ a complicated strategy to process low-quality depth information or independently process different modality data, our proposed CMFM in the first stage utilizes the self-attention mechanism for single modal features, and each of them produces an intra-modality response. Besides, we design a dynamic cross-modal guided fusion mechanism, i.e., interactive attention, to adaptively learn the weight of inter-modal information with ap-

propriate help of high-quality single modal information. Thus, CMFM can extract more useful feature representation from different modality information despite the imperfect quality of modality data. Furthermore, considering the inconsistency of cross-modal data, the useful features from the first stage are combined in an adversarial manner in the second stage. Instead of directly concatenating or adding features to generate redundancy, the adversarial combination can effectively enhance the salient areas and suppress the inconsistency in different modalities. More importantly, it can further alleviate the negative influence of unreliable complementary modality data in multi-modal SOD tasks. Moreover, we feed the efficient cross-modal fusion features generated by CMFM to the subsequent BMD to solve the issues for differently-scaled objects. BMD is dedicated to aggregate multi-scale hierarchical features from multiple levels, and its layers are constructed by a bi-directional method. Hence, BMD can capture both local and global information of salient objects to further improve the RGB-D and RGB-T SOD performance. In Fig. 1, our proposed MMNet is more effective for unreliable complementary modality data, challenging scenarios, and objects with various sizes.

The main contributions of this paper are as follows.

(1) This work is the first to investigate a unified end-to-end framework for both RGB-D and RGB-T SOD, where these two different problems are jointly analyzed and solved. To this end, we propose a unified multi-stage and multi-scale information fusion network (MMNet) for diverse multi-modal SOD tasks.

(2) We design a novel cross-modal multi-stage fusion module (CMFM) inspired by the visual color stage doctrine in the HVS. By using self-attention, interactive attention and adversarial combination, our proposed CMFM can explore the effectiveness, complementarity, and accuracy of different cross-modal features.

(3) Instead of adopting short-connections or deep supervision, we introduce a bi-directional multi-scale decoder (BMD) to reasonably associate multi-scale hierarchical cross-modal features through a bidirectional fusion structure. Specifically, the ablation experimental results on BMD demonstrate the effectiveness of aggregating multi-level multi-modal features.

(4) We present a comprehensive comparison (with $\sim 92K$ large-scale image-pairs) of 21 SOTA RGB-D and RGB-T algorithms [17]–[20], [22], [24]–[29], [44]–[53], making our work a good multi-modal SOD benchmark. Without any use of pre-processing and post-processing techniques, MMNet significantly outperforms the SOTA methods on nine challenging datasets. To the best of our knowledge, such a general framework is the first to show its robustness and generality on RGB-D and RGB-T multi-modal SOD tasks.

This work extends our previous paper [1] in several aspects. First, we offer a more in-depth discussion between visual color stage doctrine and our proposed cross-modal multi-stage fusion module (CMFM). Second, we propose a dynamic cross-modal guided fusion scheme (DCG) to further improve the performance compared with the original cross-modal guided attention module (CGA). Besides, we offer more details regarding our proposed algorithm. Third, in addition to RGB-D SOD, we provide a comprehensive analysis of RGB-T SOD

on three datasets with extensive quantitative and qualitative experiments, making our work a multi-modal benchmark for simultaneously analyzing these two problems. Fourth, we perform more ablation studies for thorough examination and provide more analyses of the experimental results.

II. RELATED WORK

Early salient object detection methods [42], [54]–[57] mainly focused on intrinsic cues or prior knowledge, such as color, texture, contrast and background. Currently, deep learning-based SOD models [7]–[15], [58] have outperformed the conventional methods in both accuracy and efficiency. Especially, fully convolutional networks (FCN) have shown effective feature generation and end-to-end training, and thus become the mainstream in SOD tasks. Considering the boundary and edge cue in saliency detection, Qin et al. [9] proposed a hybrid loss to focus on the boundary cue of salient objects, and Zhao et al. [10] utilized salient edge information within the network to preserve edge cue. Moreover, attention mechanism has been successfully applied to improve the performance of SOD. Zhao et al. [8] designed a pyramid feature attention network to explore channel and spatial attention in different levels. Liu et al. [7] utilized pixel-wise contextual attention to select useful context information. In addition, many works aimed to fuse multi-level features to achieve better results. Wei et al. [11] proposed a cross-feature module and the cascaded decoder to fuse and refine features of different levels. Li et al. [12] designed a cross-layer feature aggregation module and the distribution module to facilitate fusion between multi-level features. These RGB salient object detection methods can be further improved by introducing complementary modalities, including depth and thermal information.

A. RGB-D Salient Object Detection

Hand-Crafted based Models. The traditional RGB-D SOD research was pioneered by Niu [59] et al., who used parallax contrast to measure stereo saliency. Afterwards Peng et al. [60] proposed a multi-context RGB-D saliency detection method that combines the comparison of low-level features, such as depth and color information. Ren et al. [61] directly integrated the background prior, normalized depth prior, and surfaced orientation prior with the image area contrast. Feng et al. [51] utilized local background and foreground features, depth, and spatial priors for salient object detection. However, these simple operations mainly rely on handcrafted features and prior knowledge of the existing datasets, resulting in limited capability to detect salient objects in complex scenes.

Deep Learning based Methods. RGB-D SOD methods based on deep learning have become the mainstream trend and achieved better performances due to the powerful learning capacity of convolutional neural networks (CNNs). Qu et al. [48] utilized low-level saliency cues by manually devising CNN networks for cross-modal features integration. Chen et al. [20] designed an asymptotic network that exploits cross-modal complementarity in a top-down manner. A fusion network of cross-modalities was proposed in [46] through a multi-scale and multi-path approach. A recent work [19] proposed

a three-stream fusion network structure that extracts RGB-D feature expression in a bottom-up and top-down manner. Zhao et al. [18] enhanced depth feature expression with contrast-enhanced depth maps and used fluid pyramids to exploit multi-scale features. Piao et al. [17] proposed a depth-induced loop network that refines multi-scale fusion features by using recurrent attention modules. Fan et al. [24] tried to address the depth map quality issue by designing the depth depurator unit (DDU) to automatically discard low-quality depth maps. These methods either simply integrated RGB and depth information, which often ignored the impact of unreliable or ambiguous depth maps, or directly abandoned low-quality depth maps and only relied on RGB information for challenging scenes.

Recently, Fu et al. [62] firstly converted the depth map to a three-channel map and employed a shared backbone for six-channel RGB-D feature extraction. Zhang et al. [63] proposed a probabilistic RGB-D SOD method by applying conditional variational autoencoders for better SOD performance. Li et al. [64] captured complementarity features between RGB and depth cues in an interweaved manner with attention mechanism. Liu et al. [45] designed an attention learning way for cross-modal fusion and selective attention for final saliency detection. Zhang et al. [44] designed a complementarity interaction module (CIM) and used compensation-aware loss to improve SOD performance. Zhang et al. [65] designed an asymmetric two-stream framework consisting of a flow ladder module for fully extracting RGB information and depth attention module for effectively utilizing depth features, respectively. Liu et al. [66] obtained saliency map from different layers by selectively fusing cross-modal features and using a fusion refinement module.

B. RGB-T Salient Object Detection

For adverse environments, such as fog, haze, low illumination, etc., thermal image data could conquer the above challenges well due to the insensitivity to illumination conditions [67]. Recently, more and more works focused on integrating RGB and thermal infrared data together to exploit saliency detection. As the original research, Wang et al. [25] built up the first RGB-T SOD dataset and designed a multi-task manifold ranking algorithm for salient object detection. Later, Tu et al. [26] utilized a multi-scale cross-modal manifold ranking to integrate different features, and introduced a weight for each modality to solve the ranking model. Tu et al. [27] proposed a graph learning method, which utilized superpixels and hierarchical deep features to jointly learn graph structure, edge weights and node weights. These methods are based on traditional machine learning technologies and may not take full advantage of rich contextual information between modalities. Two recent end-to-end works have explored multi-interactive encoder-decoder [28], and multi-level features with attention mechanism [29], respectively, and achieve good performances.

Although several related works exist, it still remains an open problem that how to effectively exploit the complementarity and facilitate a deeper fusion of different modalities, i.e., RGB, depth and thermal data. To fully take advantage of cross-modalities, our proposed method considers the reliability and

importance between different modalities, and designs a multi-stage and multi-scale strategy to explore a more accurate feature representation for cross-modal fusion.

III. PROPOSED METHOD

In this section, we first introduce the model architecture of our proposed MMNet. Then, we give a detailed explanation for the correspondence between our model and the visual color stage doctrine in the HVS. Finally, we present our proposed cross-modal multi-stage fusion module (CMFM) and the bi-directional multi-scale decoder (BMD).

A. Model Architecture

As shown in Fig. 2, our proposed framework follows an encoder-decoder architecture and mainly consists of a cross-modal multi-stage fusion module (CMFM) and a bi-directional multi-scale decoder (BMD). Specifically, RGB and complementary streams (i.e., depth or thermal) have the same structure for extracting different modality features. The outputs of conv2, conv3, conv4, and conv5 from RGB and complementary branches are denoted as $\{f_{r_i}\}_{i=2}^5$ and $\{f_{c_i}\}_{i=2}^5$, respectively. Then, we fuse paired side-out features $\{f_{r_i}\}_{i=2}^5$ and $\{f_{c_i}\}_{i=2}^5$ to generate cross-modal fusion features $\{cm_i\}_{i=2}^5$ by employing the novel CMFM. It aims to discover the effectiveness and complementarity between different modalities through feature response stage and adversarial combination stage. Moreover, we feed the accurate cross-modal fusion features $\{cm_i\}_{i=2}^5$ to the subsequent BMD component. BMD is dedicated to progressively aggregate multi-scale hierarchical features from multiple levels, which is constructed by the bi-directional method. Finally, the output of BMD is supervised by the ground truths to generate saliency maps. In this sense, the complementarity, effectiveness, diversity in multi-modal SOD can be explored from a “feature-based” perspective in our model. Details will be given in the following sections.

B. Discussion of Visual Color Stage Doctrine

Most existing works [19], [28], [46] indiscriminately integrate multi-modal features from RGB and complementary backbones. However, as aforementioned, the generalization ability of such methods can be easily restricted by unreliable complementary modality data when it comes to low-quality depth maps or misaligned thermal maps [24], [29], [68]. In addition, direct fusion can introduce background noise due to the lack of consideration on the inconsistency of different modality information. To address the issue of unreliable modality data, a depth depurator unit was proposed to filter out low-quality depth maps in [24]. Zhao et al. [18] proposed a contrast-enhanced network for obtaining better contrast of original depth maps. However, the aforementioned methods often depended on a complex two-step strategy, which may introduce redundant computations and results in suboptimal results. To the best of our knowledge, there are no developed saliency methods considering RGB-D and RGB-T multi-modal fusion SOD tasks simultaneously.

The performance of multi-modal information fusion directly influences the result of SOD, and thus we should have accurate

and effective modeling for this fusion process. We propose to resort to the renowned theory of visual color stage doctrine to simulate and explain the complex information integration system in multi-modal SOD. Studies have shown that perceiving visual color information includes the tricolor response stage and the color perception stage. The response to spectrum is not a simple tricolor mixture but acts as a more complex system, which is known as the visual color stage doctrine [30]–[36]. As shown in Fig. 3 (a), the inputs of the visual color perception process are the basic optical signals and the output is the color. In the first stage, the three cone cells of retina (i.e., photoreceptor cells) selectively absorb different wavelengths of light radiations, and each of them produces a corresponding neural stimulus, i.e., red, green, and blue neural responses. In the second stage, nerve impulses fuse the corresponding red, green, and blue neural responses by the means of adversarial combinations, and then the fused signals are fed into cerebral cortex to generate the visual color expression.

Lots of research efforts [41]–[43], [48], [54], [60], [69], [70] have shown that the basic visual features are crucial to salient object detection. Meanwhile, the visual color stage doctrine describes how basic optical signals are processed in the human visual system. Therefore, from the perspective of human visual system, both of them are the integration process of basic visual signals and they have the similar mechanism, and the features from RGB, depth and thermal maps are also optical signals. Hence, the selection and combination steps in multi-modal SOD can be exactly consistent with the above two stages as mentioned in the doctrine theory. Inspired by the core mechanism of high-efficiency human visual color system, we propose a unified end-to-end framework for different multi-modal SOD problems. This framework aims to adaptively model the importance of different modality features, and simultaneously suppress the uncertain area due to the inconsistency of different modalities. Hence, it can generate a more accurate multi-modal feature expression. The detailed relationship between the visual color stage doctrine in the HVS and our proposed CMFM are discussed as follows.

Structure. As shown in Fig. 3, the visual color stage doctrine consists of the tricolor response stage and the color perception stage, while CMFM is made up of the feature responses stage and the adversarial combination stage correspondingly.

Correspondence analysis of Stage1. As the mentioned visual color stage doctrine, by selecting different wavelengths, the optical absorption behavior can produce red, green, and blue neural responses, respectively. Correspondingly, we employ self-attention mechanism and dynamic cross-modal guided fusion mechanism to select different useful modal features and generate three useful feature responses in CNN.

Correspondence analysis of Stage2. Then, red, green, and blue neural responses are combined as the opponent-color Y-B, R-G, and luminance L in an adversarial way. The final visual perception of color depends on not only the neural response generated by the color itself, but also the inhibitory effect of the opponent-color. Correspondingly, we utilize multiplication and pooling operation to take good advantage of different modality feature responses and adaptively model the impor-

Fig. 2. Overall architecture of the proposed multi-stage and multi-scale fusion network (MMNet). The main components are the proposed cross-modal multi-stage fusion module (CMFM) and bi-directional multi-scale decoder (BMD).

Fig. 3. The proposed cross-modal multi-stage fusion module (CMFM). (a) The mechanism of color vision stage doctrine. (b) Overall structure of CMFM with two similar steps as the color information processing.

tance of them, which generates a more accurate multi-modal feature expression.

It is important to note that the CNN features are the transition state of the basic optical signals, which can be deemed as pre-processing signals. This pre-processing can transform the basic optical signals into more representative features, i.e., CNN features, which are beneficial to salient object detection.

C. Proposed Cross-modal Multi-stage Fusion Module (CMFM)

Exploring accurate cross-modal fusion features is essential to address the challenging problems of multi-modal SOD. Inspired by the visual color stage doctrine in the HVS, we propose a novel cross-modal multi-stage fusion module to express the accurate cross-modal features.

Specifically, we use the spatial attention (SA) to highlight saliency, which effectively selects the important response of

Fig. 4. Different structure comparisons for inter-modal features fusion. D/T denote depth and thermal input features, respectively. (a) Element-wise sum strategy. (b) Cross-modal guided attention module (CGA) in [1]. (c) The proposed dynamic cross-modal guided fusion module (DCG).

intra-modal information in the first stage, and is described as:

$$\begin{cases} Fr_i = SA(W_i^r * fr_i) \otimes fr_i, \\ Fc_i = SA(W_i^c * fc_i) \otimes fc_i, \end{cases} \quad i \in \{2, 3, 4, 5\}, \quad (1)$$

where $\otimes$ denotes element-wise multiplication. $\{fr_i\}_{i=2}^5$ represents the side-out features of RGB branch. $\{fc_i\}_{i=2}^5$ represents the side-out features of depth or thermal branch as shown in Fig. 2. $\{W_i^r\}_{i=2}^5$ and $\{W_i^c\}_{i=2}^5$ mean 1×1 convolution operators to uniformly convert the channel of input to 256. Fr_i and Fc_i represent the enhanced modal features through the spatial attention SA to emphasize the useful features and

suppress the unnecessary ones.

More importantly, to effectively select reliable and important cross-modal features, we design a dynamic cross-modal guided fusion (DCG) module. Firstly, considering the impact of unreliable or ambiguous complementary modality data, it generates an effective inter-modal input (i.e., F_i^G) with appropriate guidance of enhanced RGB information (i.e., Fr_i) from previous spatial attention. Then, we adopt the pooling operation with different sizes and shapes, and the weighted fusion strategy to adaptively explore a more accurate cross-modal feature expression. The process can be described as:

$$F_i^G = \text{Conv}(\text{Cat}(Fr_i, fc_i)), \quad (2)$$

$$F_i^K = \sum_{m=1}^3 U(AP_K^m(F_i^G)), \quad (3)$$

$$F_i^S = U(AP_{1H}(F_i^G)) + U(AP_{W1}(F_i^G)), \quad (4)$$

$$Frc_i = \gamma^i CA(\text{Cat}(F_i^K, F_i^S)) + F_i^G, \quad (5)$$

where F_i^G is the guided fusion feature with the help of the enhanced RGB modality feature. Cat denotes concatenation operation, and Conv consists of a 3×3 convolution and a ReLU layer. AP represents adaptive average pooling, and then we use up-sampling U to resize to the original input size. Specifically, inspired by [71], we propose the strip pooling (i.e., AP_{1H} , AP_{W1}) and the square global average pooling (i.e., AP_K^m) to dynamically capture both local and global cross-modal context. For example, given an inter-modal feature map $x_l \in \mathbb{R}^{C \times W \times H}$, we obtain strip output with $(W, 1)$ and $(1, H)$ size. Meanwhile, we propose the 3-level square global average pooling with (K, K) output size. Here, $K \in \{1, 2, 4\}$ ($i = 4, 5$) and $K \in \{1, 4, 8\}$ ($i = 2, 3$). Then, we implement the weighted fusion strategy through channel attention (CA), and set a learnable parameter γ^i to adjust the relative ratio between input and weighted features. $\{Frc_i\}_{i=2}^5$ means the dynamic cross-modal guided fusion feature.

Since there exists the imperfectness of single modal data and the inconsistency of different feature responses generated in the first stage, simply adding may induce the suboptimal results that ignore the importance and reliability of different feature responses. Hence, an adversarial combination method is designed to take good advantage of different modality features and adaptively model the importance of them. The representation of cross-modal feature depends not only on the self-response from single modality but also on the adversarial effect from another modality. The second stage conducts the combination as:

$$A_L^i = W_L^i * (Fr_i \oplus Fc_i \oplus Frc_i), \quad (6)$$

$$A_{RC'}^i = W_{RC'}^i * (Fr_i \otimes (Fc_i \oplus Frc_i)), \quad (7)$$

$$A_{CC'}^i = W_{CC'}^i * (Fc_i \otimes Frc_i), \quad (8)$$

$$F_v^i = AP(W_v^i * \text{Cat}(A_{CC'}^i, A_{RC'}^i, A_L^i)), \quad (9)$$

$$F_m^i = MP(W_m^i * \text{Cat}(A_{CC'}^i, A_{RC'}^i, A_L^i)), \quad (10)$$

$$cm_i = \sigma(W_v^i * F_v^i \oplus W_m^i * F_m^i), \quad (11)$$

Fig. 5. Feature fusion (FF) module and multi-scale feature fusion (MF) module in the proposed bi-directional multi-scale decoder (BMD).

where $\oplus$ and $\otimes$ denote the element-wise addition and multiplication, respectively. W_* are model parameters to be learned. AP and MP mean adaptive average pooling operation and adaptive max pooling operation, respectively. σ denotes the Sigmoid operation. Then, the cross-modal fusion features $\{cm_i\}_{i=2}^5$ can be regarded as the inputs to the layers in BMD.

D. Proposed Bi-directional Multi-Scale Decoder (BMD)

Based on the insightful work [72], different layers have different responses to the entire objects. Specifically, the features in deep layers provide discriminative and comprehensive semantic information, while shallow features contain more local detail information. Because the distribution of salient objects is quite different in various scenes, we introduce BMD to aggregate multi-scale and comprehensive features from both the top-down and bottom-up pathways.

In Fig. 2, the output features $\{cm_i\}_{i=2}^5$ from CMFM contain cross-modal fusion features on each layer, which are fed to feature fusion (FF) module in a top-down pathway firstly. Then, for each multi-scale feature fusion (MF) module, the side out features $\{F_j\}_{j=2}^5$ from FF and cross-modal fusion features $\{cm_i\}_{i=2}^5$ from CMFM can be integrated in a bottom-up manner. The multi-scale cross-modal fusion features from low-level and high-level can be aggregated, and the salient objects can be highlighted with complete structure and sharp boundary. Specifically, Fig. 4 illustrates the details.

In each FF, the inputs are $\{cm_i\}_{i=2}^5$ from CMFM and FF features $\{F_{j+1}\}_{j=2}^4$ from the upper level. Therefore, the feature fusion in FF can be described as:

$$\begin{cases} f_i = (\widetilde{W}_i * cm_i) \otimes (\widetilde{W}_{j+1} * AP(F_{j+1})), \\ F_j = W_i * f_i \oplus W_{j+1} * U(F_{j+1}), \end{cases} \quad i, j \in \{2, 3, 4\}, \quad (12)$$

$$F_j = W_i * cm_i, \quad i, j = 5, \quad (13)$$

where $\oplus$ and $\otimes$ denote element-wise addition and multiplication, respectively, and U is the 2x upsampling operation via bilinear interpolation. AP means global average pooling

operation. $\{\widetilde{W}_i\}_{i=2}^4$ and $\{\widetilde{W}_{j+1}\}_{j=2}^4$ denote 3×3 convolution operation with PReLU activation and 1×1 convolution operation with ReLU activation, respectively, and they uniformly convert the channels of $\{cm_i\}_{i=2}^5$ to 64. $\{W_i\}_{i=2}^5$ and $\{W_{j+1}\}_{j=2}^5$ only mean 3×3 convolution operation without activation. The fusion result F_j is exactly the higher level feature of the next fusion stage.

Then, the integrated features $\{F_j\}_{j=2}^5$ from FF and cross-modal fusion features $\{cm_i\}_{i=2}^5$ are fed into MF, which can be described as:

$$\begin{cases} S_k = scSE(\widetilde{W}_{\sum}^k * (F_j \oplus (W_{sk}^i * cm_i) \\ \oplus (W_{\sum}^{k-1} * S_{k-1}))), i, j, k \in \{3, 4, 5\}, \\ S_k = scSE(\widetilde{W}_{\sum}^k * (F_j \oplus (W_{sk}^i * cm_i))), i, j, k = 2, \end{cases} \quad (14)$$

where $\oplus$ denotes the element-wise addition. $\{W_{sk}^i\}_{i=2}^5$ and $\{\widetilde{W}_{\sum}^k\}_{k=2}^5$ represent 1×1 convolution operations with ReLU activation. $\{W_{\sum}^{k-1}\}_{k=3}^5$ denotes the 3×3 downsampling convolution operation with stride of 2. The scSE from [73] is adopted to capture more meaningful detailed information.

Finally, the MF outputs with the finest features are denoted as $\{S_k\}_{k=2}^5$, and we upsample them to the original same size. Then, we concatenate them, and the final saliency map S_{map} can be obtained by a prediction convolution layer:

$$S_{map} = \mathbf{P}(Cat_{k=2}^5 U(S_k)), \quad (15)$$

where U is the upsampling operation, and $\mathbf{P}$ means the 3×3 prediction convolution layer.

IV. EXPERIMENTS

A. Evaluation Datasets and Metrics

1) *RGBD-based SOD Datasets*: To comprehensively evaluate RGB-D SOD performance, we conduct experiments on six public RGB-D datasets. LFS [74] contains 100 images and the depth maps are obtained by Lytro camera. NLPR [60] consists of 1000 images taken by Kinect under different illumination conditions. NJUD [50] is a big dataset, which includes 1985 images collected from 3D movies, Internet, and photographs. STERE [59] contains 1000 internet stereoscopic images and the depth maps are estimated from stereoscopic images. SSD [75] consists of 80 test images picked up from movies. DUT-RGBD [17] includes 1200 challenging images captured by Lytro camera in real-world scenes.

2) *RGBT-based SOD Datasets*: Moreover, we conduct experiments on three RGB-T SOD datasets, which further demonstrates the validity of our proposed model in multi-modal fusion task. VT821 [25] consists of 821 RGB-T image pairs taken by a thermal imager (FLIR A310) and a CCD camera (SONY TD-2073). VT1000 [27] contains 1000 RGB-T image pairs, which are captured by FLIR SC620 camera. VT5000 [29] is the latest RGB-T dataset, which includes 5000 challenging images pairs collected by FLIR T640 and T610.

3) *Evaluation Metrics*: We adopt recognized E-measure (E_{ξ}) [80], S-measure (S_m) [76], adaptive F-measure (F_{β}) [78], weighted F-measure (F_{β}^w) [81] and MAE for evaluation. E-measure jointly considers both global statistics and

local pixel matching information. S-measure means the spatial structure similarities, and combines the region-aware S_r and object-aware S_o structural similarity as:

$$S_m = \alpha * S_o + (1 - \alpha) * S_r, \quad (16)$$

where α in $[0,1]$ is the balance parameter and set to 0.5 [76].

Applying thresholding on the predicted saliency map (from 0 to 255), the pairs of precision-recall scores can be calculated by comparing the binary mask with ground truth. F-measure can evaluate the overall performance and is a harmonic mean of precision and recall, and can be formulated as:

$$F_{\beta} = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \quad (17)$$

where β^2 is empirically set 0.3 to give more emphasis on precision than recall as suggested in [78]. Weighted F-measure modulates F-measure by weighting the Precision and Recall, which assigns different errors at different locations. The MAE indicates the average absolute difference between saliency map S and ground truth G , and is computed by:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |S(x, y) - G(x, y)|, \quad (18)$$

where W and H represent image width and height. Higher S_m , F_{β} , F_{β}^w , E_{ξ} and lower MAE denote better saliency results.

B. Implementation Details

We conduct our experiments based on PyTorch with a single NVIDIA GeForce RTX 2080Ti GPU. The training and testing images are uniformly resized to 256×256 . In this work, the parameters of the backbone network are initialized by VGG19 [82] or Res2Net-50 [83] and other convolutional parameters are randomly initialized. The ablation studies are conducted under Res2Net-50 backbone and models are trained with RGB-D data. For RGBD-based SOD task, we adopt the same training set [17], [44], [84] and augment the training set by cropping, flipping, and rotating. The remaining images and datasets are all used for testing. We use the stochastic gradient descent (SGD) method and cross-entropy loss function to train the network, where the initial learning rate, weight decay, and momentum are set to $5e-9$, 0.0005, and 0.99, respectively. The learning rate is divided by 10 after 35 epochs. For RGBT-based SOD task, following [29], we use 2500 samples as the training set, and the remaining samples and datasets are used for testing. Training data are only augmented by flipping. Adam algorithm is adopted to optimize our model. The initial learning rate is set as $1e-5$ and is divided by 10 after 20 epochs.

C. Performance of RGB-D Salient Object Detection

1) *Compared Methods*: The proposed method is compared with 16 RGB-D salient object detection methods, including 11 deep learning-based methods: SSF [44], S2MA [45], D3Net [24], DMRA [17], CPFP [18], TANet [19], PCF [20], MMCI [46], CTMF [47], AFNet [22], DF [48]. In addition, we also report the results of traditional RGB-D SOD works, including SE [49], ACSD [50], LBE [51], DCMC [52], MDSF [53].

TABLE I

QUANTITATIVE COMPARISON OF S-MEASURE (S_m) [76], MEAN ABSOLUTE ERROR (MAE) [77], ADAPTIVE F-MEASURE (F_β) [78], [79], ADAPTIVE E-MEASURE (E_ξ) [80] ON FIVE WIDELY-USED RGB-D DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. $\uparrow$ AND $\downarrow$ DENOTE THAT LARGER AND SMALLER ARE BETTER, RESPECTIVELY. OUR METHOD RANKS THE FIRST ON ALL DATASETS AND ALL METRICS.

Dataset	Metric	ACSD [50]	LBE [51]	DCMC [52]	MDSF [53]	SE [49]	DF [48]	AFNet [22]	CTMF [47]	MMCI [46]	PCF [20]	TANet [19]	CPFP [18]	DMRA [17]	D3Net [24]	S2MA [45]	SSF [44]	MMNet VGG19	MMNet Res2Net50
NJUD [50]	$S_m \uparrow$	0.699	0.695	0.686	0.748	0.664	0.763	0.772	0.849	0.858	0.877	0.878	0.878	0.886	0.895	0.894	0.899	0.911	0.910
	$MAE \downarrow$	0.202	0.153	0.172	0.157	0.169	0.141	0.100	0.085	0.079	0.059	0.060	0.053	0.051	0.051	0.053	0.043	0.038	0.038
	$F_\beta \uparrow$	0.696	0.740	0.717	0.757	0.734	0.784	0.768	0.788	0.812	0.844	0.844	0.837	0.872	0.840	0.865	0.886	0.901	0.899
	$E_\xi \uparrow$	0.786	0.791	0.791	0.812	0.772	0.835	0.846	0.864	0.878	0.896	0.893	0.895	0.908	0.892	0.896	0.912	0.920	0.922
NIPR [60]	$S_m \uparrow$	0.673	0.762	0.724	0.805	0.756	0.802	0.799	0.860	0.856	0.874	0.886	0.888	0.899	0.906	0.915	0.914	0.923	0.925
	$MAE \downarrow$	0.179	0.081	0.117	0.095	0.091	0.085	0.058	0.056	0.059	0.044	0.041	0.036	0.031	0.034	0.030	0.026	0.024	0.023
	$F_\beta \uparrow$	0.535	0.736	0.614	0.665	0.692	0.744	0.747	0.724	0.730	0.795	0.796	0.823	0.855	0.834	0.853	0.875	0.892	0.892
	$E_\xi \uparrow$	0.742	0.855	0.786	0.812	0.839	0.868	0.884	0.869	0.872	0.916	0.916	0.924	0.942	0.932	0.938	0.949	0.951	0.955
LFSD [74]	$S_m \uparrow$	0.734	0.736	0.753	0.700	0.698	0.791	0.738	0.796	0.787	0.794	0.801	0.828	0.847	0.832	0.837	0.859	0.866	0.878
	$MAE \downarrow$	0.188	0.208	0.155	0.190	0.167	0.138	0.133	0.119	0.132	0.112	0.111	0.088	0.075	0.099	0.094	0.066	0.065	0.057
	$F_\beta \uparrow$	0.755	0.708	0.816	0.799	0.778	0.806	0.742	0.782	0.779	0.792	0.794	0.813	0.849	0.801	0.820	0.867	0.864	0.875
	$E_\xi \uparrow$	0.801	0.770	0.842	0.817	0.784	0.844	0.810	0.851	0.840	0.842	0.845	0.867	0.883	0.833	0.863	0.895	0.896	0.913
STERE [59]	$S_m \uparrow$	0.692	0.660	0.731	0.728	0.708	0.757	0.825	0.848	0.873	0.875	0.871	0.879	0.835	0.889	0.890	0.837	0.890	0.884
	$MAE \downarrow$	0.200	0.250	0.148	0.176	0.143	0.141	0.075	0.086	0.068	0.064	0.060	0.051	0.066	0.054	0.051	0.065	0.044	0.046
	$F_\beta \uparrow$	0.661	0.595	0.742	0.744	0.748	0.742	0.807	0.771	0.829	0.826	0.835	0.830	0.844	0.833	0.855	0.841	0.881	0.868
	$E_\xi \uparrow$	0.793	0.749	0.831	0.830	0.825	0.838	0.886	0.864	0.901	0.897	0.906	0.903	0.900	0.904	0.907	0.899	0.925	0.916
SSD [75]	$S_m \uparrow$	0.675	0.621	0.704	0.673	0.675	0.747	0.714	0.776	0.813	0.841	0.839	0.807	0.857	0.866	0.868	0.844	0.877	0.871
	$MAE \downarrow$	0.203	0.278	0.169	0.192	0.165	0.142	0.118	0.099	0.082	0.062	0.063	0.082	0.058	0.058	0.052	0.058	0.044	0.047
	$F_\beta \uparrow$	0.656	0.613	0.679	0.674	0.693	0.724	0.694	0.710	0.748	0.791	0.767	0.726	0.821	0.793	0.818	0.809	0.845	0.832
	$E_\xi \uparrow$	0.765	0.729	0.786	0.772	0.778	0.812	0.803	0.838	0.860	0.886	0.879	0.832	0.892	0.888	0.891	0.899	0.922	0.903
DUT [17]	$S_m \uparrow$	0.361	—	0.499	—	—	0.730	—	0.831	—	—	0.808	0.749	0.888	—	0.903	0.915	0.919	0.920
	$MAE \downarrow$	0.332	—	0.243	—	—	0.145	—	0.097	—	—	0.093	0.099	0.048	—	0.043	0.033	0.032	0.032
	$F_\beta \uparrow$	0.128	—	0.406	—	—	0.748	—	0.792	—	—	0.779	0.736	0.883	—	0.886	0.915	0.924	0.918
	$E_\xi \uparrow$	0.621	—	0.712	—	—	0.842	—	0.882	—	—	0.866	0.814	0.927	—	0.921	0.946	0.951	0.951

Fig. 6. Qualitative comparison of the proposed MMNet (Res2Net) and the recent competitive RGB-D methods. (a) RGB image. (b) Depth map. (c) Ground truth. (d) Our MMNet. (e) SSF [44]. (f) S2MA [45]. (g) D3Net [24]. (h) DMRA [17]. (i) TANet [19]. (j) CPFP [18]. (k) PCF [20]. (l) AFNet [22]. (m) SE [49]. The inputs show various challenging scenes, including low-quality or confusing depth maps (rows 1-3), complex background (row 4), low contrast (row 5), and multi-scale salient objects (rows 6-8).

2) *Quantitative Comparisons:* From Table I, our model significantly outperforms the recently proposed methods, such as S2MA [45], SSF [44], D3Net [24], DMRA [17], and the notable improvements can be consistently observed in all

TABLE II

QUANTITATIVE COMPARISON OF S-MEASURE (S_m) [76], MEAN ABSOLUTE ERROR (MAE) [77], ADAPTIVE F-MEASURE (F_β) [78], WEIGHTED F-MEASURE (F_β^w) [81], E-MEASURE (E_ξ) [80] ON THREE WIDELY-USED RGB-T DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. $\uparrow$ AND $\downarrow$ DENOTE THAT LARGER AND SMALLER ARE BETTER, RESPECTIVELY. OUR METHOD RANKS THE FIRST ON ALL DATASETS AND ALL METRICS.

Method	VT821 Dataset [25]					VT1000 Dataset [27]					VT5000 Dataset [29]				
	$E_\xi \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\xi \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$E_\xi \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$
MTMR [25]	0.815	0.725	0.462	0.662	0.108	0.836	0.706	0.485	0.715	0.119	0.795	0.680	0.397	0.595	0.114
M3S-NIR [26]	0.859	0.723	0.407	0.734	0.140	0.827	0.726	0.463	0.717	0.145	0.780	0.652	0.327	0.575	0.168
SGDL [27]	0.847	0.765	0.583	0.730	0.085	0.856	0.787	0.652	0.764	0.090	0.824	0.750	0.558	0.672	0.089
MIED [28]	0.872	0.866	0.705	0.765	0.055	—	—	—	—	—	0.842	0.829	0.659	0.709	0.064
ADNet [29]	0.842	0.810	0.626	0.716	0.077	0.921	0.910	0.804	0.847	0.034	0.891	0.863	0.730	0.777	0.048
DMRA* [17]	0.878	0.791	0.689	0.760	0.056	0.887	0.836	0.770	0.809	0.055	0.857	0.771	0.652	0.704	0.062
Ours(Res2Net50)	0.892	0.873	0.793	0.795	0.040	0.923	0.913	0.869	0.863	0.027	0.888	0.862	0.779	0.781	0.043
Ours*(VGG19)	0.919	0.883	0.832	0.843	0.033	0.926	0.905	0.876	0.874	0.032	0.896	0.850	0.783	0.796	0.045
Ours(VGG19)	0.894	0.874	0.791	0.799	0.040	0.924	0.917	0.865	0.865	0.027	0.890	0.863	0.782	0.786	0.043

TABLE III

RESULTS FOR ABLATION STUDIES WITH DIFFERENT COMBINATIONS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

#	Component					NJUD [50]		LFSD [74]		VT821 [25]		VT1000 [27]	
	Depth	Stage1	Stage2	FPN	BMD	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$
1	✓			✓		0.853	0.054	0.839	0.082	0.759	0.042	0.807	0.040
2	✓	✓		✓		0.876	0.045	0.857	0.069	0.800	0.037	0.838	0.033
3	✓	✓			✓	0.889	0.041	0.862	0.066	0.813	0.035	0.854	0.030
4	✓	✓	✓	✓		0.884	0.043	0.868	0.063	0.810	0.037	0.855	0.032
5	✓	✓	✓		✓	0.899	0.038	0.875	0.057	0.830	0.033	0.865	0.028

metrics. Especially, from the proposed cross-modal multi-stage fusion module (CMFM), our method achieves impressive gains on NLPR [60] and DUT-RGBD [17], which contain more complicated scenes. This fact further demonstrates the consistent effectiveness and robustness of our model. Moreover, the SSD [75] dataset has lots of confusing depth maps. Our MMNet can obtain performance gains with $\sim 5\%$ improvement on F-measure score and 24.1% reduction on MAE score over D3Net [24] that is originally designed for addressing the depth map quality issue. We attribute this gain to its powerful ability in utilizing depth information more effectively by CMFM. The proposed MMNet also performs the best on LFSD [74] and NJUD [50] datasets, which contain various multi-scale salient objects [24]. It also validates that our model is more powerful in locating salient objects with different sizes.

3) *Visual Comparisons*: In Fig. 6, we show some visualized examples to further demonstrate the superiority of our MMNet by comparisons with other recent competitive methods. It can be seen that our obtained saliency results are much closer to the ground truths. As shown in rows 1-3, the existing RGB-D methods cannot accurately detect the salient regions due to the confusing depth maps. On the contrary, the proposed MMNet can achieve better performances with highlighted foregrounds and complete structures. Moreover, rows 4-5 show different challenging scenes, including complex backgrounds and low contrast scenes. Our method effectively utilizes cross-modal complementary information for capturing salient regions. Additionally, due to the proposed bi-directional multi-scale decoder (BMD), our MMNet not only locates the salient objects with different sizes but also precisely distinguishes complete structures and coherent details as shown in rows 6-8.

D. Performance of RGB-T Salient Object Detection

1) *Compared Methods*: The 6 different state-of-the-art multi-modal SOD methods are compared on three challenging

datasets, including 5 RGB-T SOD methods: ADNet [29], MIED [28], SGDL [27], M3S-NIR [26], MTMR [25]. Furthermore, we report a RGB-D SOD method DMRA* [17] for reference, which is trained with RGB-D data. For fair comparisons, we use their published saliency maps or released source codes to generate relevant results in testing.

2) *Quantitative Comparisons*: In Table II, it can be observed that our proposed method consistently outperforms the existing methods in terms of all testing datasets on E_m , S_m , F_m , F_β^w and MAE values. Especially on the challenging VT821 dataset, which has some alignment errors in RGB-T pairs, 8.6% and 3.4% gains on F_β^w and F_m scores can be seen when compared with the second best, respectively. Similar results can also be seen on the other results. This indicates that CMFM can adaptively select effective modal features and utilize complementary information within RGB-T pairs. For VT1000 and VT5000 datasets that have more complex and diversified scenes, our method also achieves 6.1% and 5.2% improvements on F_β^w scores, respectively. It also demonstrates the effectiveness of our proposed method for complex and diversified multi-modal scenes. In addition, our MMNet* (VGG19) still achieves better saliency results than other state-of-the-art RGB-T methods on all datasets, which is trained on RGB-D data instead of RGB-T data. This also indicates that our proposed method possesses the good generalization capability in the multi-modal fusion tasks.

3) *Visual Comparisons*: To qualitatively evaluate our proposed method for RGB-T scenes, various representative samples are visualized in Fig. 7, including image clutter, out of focus, thermal crossover, similar appearance, low illumination, multiple salient objects, big salient objects, and small salient objects. It can be also observed that our saliency maps are closer to the ground truths.

Complex Scenes. For complex or cluttered backgrounds shown in rows 1-2 of Fig. 7, our method can achieve better per-

Image clutter

Out of focus

Thermal crossover

Similar appearance

Low illumination

Multiple salient objects

Big salient objects

Small salient objects

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 7. Qualitative comparison of the proposed MMNet (Res2Net) and the recent competitive RGB-T methods. (a) RGB image. (b) Thermal image. (c) Ground truth. (d) Our MMNet. (e) ADFNet [29]. (f) DMRA* [17]. (g) MIED [28]. (h) SGDL [27]. (i) M3S-NIR [26]. (j) MTMR [25]. The inputs show various challenging scenes, including image clutter (rows 1-2), out of focus (rows 3-4), thermal crossover (rows 5-6), similar appearance (rows 7-8), low illumination (rows 9-10), multiple salient objects (rows 11-12), big salient objects (rows 13-14) and small salient objects (rows 15-16).

formances in terms of completeness while the existing methods cannot address this challenging case very well. Besides, the proposed method can effectively utilize complementary cross-modal features to capture accurate salient objects when the object is out-of-focus or blurred in rows 3-4.

Low Contrast. From rows 5-8 of Fig. 7, we can see that the salient objects have a similar color with the background in RGB or infrared thermal image, and thus many methods cannot accurately detect them. In contrast, our model can make use of complementary thermal information to improve

Fig. 8. The baseline of our proposed network. $\oplus$ denotes element-wise addition operation.

TABLE IV

COMPARISON BETWEEN DSS [85] AND THE DESIGNED BMD. THE * DENOTES THAT THE METHOD EMPLOYS THE CRF [86] POST-PROCESSING TECHNIQUE. M DENOTES THE MAE METRIC.

Settings	NJUD [50]			VT821 [25]			VT5000 [29]		
	$S_m \uparrow$	$F_\beta \uparrow$	M $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	M $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	M $\downarrow$
DSS	0.879	0.852	0.056	0.826	0.740	0.049	0.780	0.688	0.059
DSS*	0.874	0.876	0.048	0.819	0.764	0.045	0.773	0.718	0.055
BMD	0.905	0.877	0.044	0.873	0.783	0.036	0.840	0.729	0.050
BMD*	0.902	0.896	0.040	0.872	0.832	0.032	0.842	0.795	0.045

TABLE V

EFFECTIVENESS ANALYSIS OF THE CONNECTIONS IN STAGE2.

Strategies	NJUD [50]		LFSD [74]		VT1000 [27]	
	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$
No weighting	0.876	0.045	0.857	0.069	0.838	0.033
Directly weighting	0.879	0.044	0.855	0.071	0.845	0.032
Stage2 connection	0.884	0.043	0.868	0.063	0.855	0.032

saliency results with clean background. Moreover, for those images with poor illumination and small sizes in rows 9-10, the proposed method can still successfully obtain much clearer boundaries with conspicuous regions.

Multi-scale Salient Objects. From the last six rows of Fig. 7, for those objects with different sizes and different numbers, most existing methods easily miss accurate salient regions or contain irrelevant non-salient regions. Differently, our proposed method not only clearly locates the salient objects with various sizes but also filters out the redundant information and obtains the complete contours. The superior results in these scenes illustrate that our proposed method can more effectively associate hierarchical features for multi-scale salient object inference.

E. Ablation Studies and Discussion

In this section, we analyze each key component of our designed model firstly. Then, we investigate the effectiveness of the proposed CMFM by removing or replacing different components. Finally, we perform the impact of different backbones to our MMNet for multi-modal SOD tasks.

1) *Effectiveness of Different Component:* The baseline model contains the same backbones and multi-layer integration structure as shown in Fig. 8. We gradually add different components to the baseline, including depth information (Depth), Stage1 with element-wise addition output in each module (Stage1), Stage2 as discussed in CMFM (Stage2), and bi-directional multi-scale decoder (BMD). Especially, our proposed model is only trained with RGB-D data and the FPN [87] structure is applied as a decoder for comparison.

Fig. 9. Visual comparison in the ablation studies. D/T represent depth and thermal data. ‘#’ denotes the corresponding row of Table III. #1: Dep + FPN. #2: Dep + Stage1 + FPN. #3: Dep + Stage1 + BMD. #4: Dep + Stage1 + Stage2 + FPN. #5: Dep + Stage1 + Stage2 + BMD.

Effectiveness of CMFM. From Table III and Fig. 9, ‘+Stage 1’ consistently achieves impressive gains. By applying Stage1, it can select useful cross-modal information, and salient objects are highlighted from Fig. 9. However, without a deeper understanding on the differences between multi-modal features, it may introduce inconsistent noises between cross-modalities during direct fusion, such as the pillar behind the bike in the second row. Similar better saliency results are shown in ‘#4’ when applying Stage2. It excels at suppressing confusing features between different modalities and can substantially locate the salient objects with less noises and more complete structures. For example, we can find that the inside of salient object is more accurate after applying Stage2 as shown in the last row of Fig. 9. In general, the quantitative results and visualizations of employing our CMFM illustrate the advantages of utilizing cross-modal information.

Effectiveness of BMD. One of our core claims is that aggregating multi-scale and multi-level features can further benefit accurate salient object detection. To give evidence for this claim, we add the BMD module for evaluation. From Table III, we obtain impressive accuracy gains after applying BMD. Furthermore, employing BMD can help to capture more details as shown in ‘#3’ of Fig. 9. However, it indiscriminately integrates different modality features from Stage1 without the adversarial combination fusion in Stage2, which sometimes leads to defective saliency results with inconsistent noise. When applying BMD together with CMFM, salient objects can show more complete structures and coherent details as ‘#5’. These demonstrate the advantage of our BMD in sufficiently associating multi-scale hierarchical cross-modal features.

Besides, to better show the strength of BMD, we compare it with DSS [85] in Table IV, which utilizes a unidirectional skip-layer structure with short connections and deep supervision for saliency detection. Specifically, based on the baseline shown in Fig. 8, we employ DSS and BMD as decoder, respectively. The results in Table IV show that our BMD is clearly superior to DSS, even DSS*, which uses CRF [86]. Moreover, the S-measure score demonstrates that salient objects show more complete structures by applying BMD. It further shows its effectiveness of aggregating multi-level features.

2) *Discussion of Cross-modal Multi-stage Fusion Module:* **Effectiveness of Connections in Stage2.** To investigate whether the connections in Stage2 play a critical role in multi-modal fusion, we implement different combination strategies

TABLE VI
ABLATION ANALYSIS OF CMFM WITH DIFFERENT DISTRIBUTION SETTINGS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Settings	NLPR [60]		NJUD [50]		LFSD [74]		VT821 [25]		VT5000 [29]	
	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$
Simple Fusion	0.853	0.028	0.874	0.045	0.860	0.066	0.790	0.036	0.742	0.049
{CM ₅ }	0.869	0.027	0.884	0.043	0.864	0.064	0.809	0.034	0.752	0.049
{CM ₅ , CM ₄ }	0.877	0.026	0.893	0.040	0.870	0.061	0.814	0.034	0.758	0.048
{CM ₅ , CM ₄ , CM ₃ }	0.885	0.024	0.895	0.039	0.871	0.062	0.821	0.033	0.771	0.047
{CM ₅ , CM ₄ , CM ₃ , CM ₂ }	0.892	0.023	0.899	0.038	0.875	0.057	0.830	0.033	0.777	0.045

TABLE VII
QUANTITATIVE COMPARISON ON THE NUMBER OF FILTERS IN CMFM.
THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Number of Filters	NLPR [60]		NJUD [50]		VT821 [25]	
	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$
128-256-256-512	0.878	0.025	0.891	0.040	0.816	0.037
128-128-128-256	0.877	0.025	0.893	0.040	0.815	0.034
128-128-256-256	0.880	0.025	0.894	0.040	0.818	0.033
128-128-256-512	0.886	0.024	0.895	0.039	0.823	0.034
256-256-256-512	0.884	0.024	0.893	0.040	0.829	0.035
256-256-256-256	0.892	0.023	0.899	0.038	0.830	0.033

TABLE VIII
QUANTITATIVE COMPARISON ON DIFFERENT INPUT SETTINGS OF DCG.
THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Settings	DUT [17]		NLPR [60]		VT821 [25]	
	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$
fr & fc	0.905	0.039	0.865	0.031	0.815	0.034
fr & Fc	0.904	0.038	0.863	0.031	0.814	0.036
Fr & Fc	0.909	0.038	0.865	0.032	0.818	0.034
Fr & fc	0.918	0.032	0.892	0.023	0.830	0.033

Fig. 10. Visual comparisons of different Stage2 variants. (a) RGB image. (b) Depth map or thermal image. (c) The saliency results of no weighting. (d) The saliency results of directly weighting. (e) The saliency results of Stage2 connection. (f) GT.

for comparisons, including:

(a) **No Weighting.** The feature maps from Stage1 at each level are directly added, followed by a 3×3 convolution layer.

(b) **Directly Weighting.** The feature maps from Stage1 are firstly added. Then, the fusion feature map is directly weighted by max and mean average pooling and sigmoid operation.

(c) **Stage2 Strategy.** We apply the proposed adversarial combination method to learn the complementarity and suppress the inconsistency between different feature responses from Stage1, as discussed in Section III-C.

Table V shows the qualitative saliency results of the above configurations under FPN decoder structure. It can be seen that both settings (b) and (c) exceed setting (a) in NJUD and VT1000 datasets. This indicates that simply adding induces suboptimal saliency results, which neglects the importance of different feature responses from Stage 1. However, the directly weighting operation may ignore the interdependencies among different inputs, which shows degraded saliency results

in LFSD dataset. Besides, the performance of setting (b) is inferior to setting (c). The reason is that the adversarial combination in Stage2 considers the existing inconsistency of different feature maps and combines each feature response interactively. In Fig. 10, the visualized results are in line with the quantitative results in Table V. Therefore, we can see that feature maps after adversarial combination are closer to ground truths with less background noises and higher confidence.

Effectiveness of CMFM with Different Levels. In order to verify that our proposed CMFM is effective at different layers, another ablation study is conducted under BMD decoder structure. Specifically, “Simple Fusion” refers to a general convolution cross-modal fusion strategy. Then, we gradually replace each general convolution layer with CMFM, while keeping other settings unchanged. As shown in Table VI, it can be observed that CMFM in each layer contributes to the saliency results on higher F-measure and lower MAE values, and thus each CMFM plays an important role for cross-modal features fusion. Moreover, we can find that CMFM applied in higher layers achieves higher gains for most datasets. The underlying reason is that the deeper layers may have similar semantic information between cross-modal features, which contributes more to the final results.

Effectiveness of CMFM with Different Channels. To investigate the number of filters in CMFM, we provide several versions of our proposed CMFM. From quantitative comparisons in Table VII, a bigger number of filters can boost performance to some extent. In general, the configuration of filter numbers {256, 256, 256, 256} achieves a better saliency result. Hence, we choose this configuration as our final setting.

Effectiveness of DCG Modules. To have a deeper understanding of the input settings in DCG modules, we report the quantitative results when using different input settings of DCG modules in Table VIII. We can see that equally combining RGB modality data and complementary modality data, or simply enhancing complementary modality features may obtain suboptimal saliency results. On the contrary, with appropriate guidance of enhanced RGB features, the combination of fc and

TABLE IX
EFFECTIVENESS ANALYSIS OF DYNAMIC CROSS-MODAL GUIDED FUSION MODULE (DCG). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Strategies	NLPR [60]		NJUD [50]		VT821 [25]		VT1000 [27]	
	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$
Element-wise sum	0.878	0.026	0.889	0.041	0.813	0.034	0.851	0.031
CGA module	0.885	0.024	0.897	0.039	0.820	0.034	0.856	0.030
DCG module	0.892	0.023	0.899	0.038	0.830	0.033	0.865	0.028

TABLE X
PERFORMANCE COMPARISON UNDER DIFFERENT BACKBONES.

Backbone	DMRA	VGG19	VGG19*	ResNet50	Res2Net50
FPS	20	30	29	23	21
NLPR	$S_m \uparrow$	0.899	0.923	0.923	0.925
	$E_m \uparrow$	0.942	0.951	0.953	0.955
	$F_m \uparrow$	0.855	0.892	0.889	0.892
	MAE $\downarrow$	0.031	0.024	0.024	0.023
NJUD	$S_m \uparrow$	0.886	0.911	0.912	0.909
	$E_m \uparrow$	0.908	0.920	0.923	0.921
	$F_m \uparrow$	0.872	0.901	0.900	0.894
	MAE $\downarrow$	0.051	0.038	0.038	0.040
LFSD	$S_m \uparrow$	0.847	0.866	0.875	0.881
	$E_m \uparrow$	0.883	0.896	0.900	0.911
	$F_m \uparrow$	0.849	0.864	0.869	0.875
	MAE $\downarrow$	0.075	0.065	0.061	0.058
STERE	$S_m \uparrow$	0.835	0.890	0.887	0.896
	$E_m \uparrow$	0.900	0.925	0.920	0.925
	$F_m \uparrow$	0.844	0.881	0.876	0.878
	MAE $\downarrow$	0.066	0.044	0.046	0.043
DUT	$S_m \uparrow$	0.888	0.919	0.922	0.919
	$E_m \uparrow$	0.927	0.951	0.949	0.946
	$F_m \uparrow$	0.883	0.924	0.921	0.916
	MAE $\downarrow$	0.048	0.032	0.031	0.032
VT821	$S_m \uparrow$	0.791	0.883	0.879	0.893
	$E_m \uparrow$	0.878	0.919	0.914	0.918
	$F_m \uparrow$	0.760	0.843	0.833	0.835
	MAE $\downarrow$	0.056	0.033	0.034	0.030
VT1000	$S_m \uparrow$	0.836	0.905	0.909	0.913
	$E_m \uparrow$	0.887	0.926	0.929	0.928
	$F_m \uparrow$	0.809	0.874	0.874	0.868
	MAE $\downarrow$	0.055	0.032	0.029	0.027
VT5000	$S_m \uparrow$	0.771	0.850	0.852	0.856
	$E_m \uparrow$	0.857	0.896	0.896	0.894
	$F_m \uparrow$	0.704	0.796	0.791	0.787
	MAE $\downarrow$	0.062	0.045	0.045	0.042

Fr can achieve better performance. It further demonstrates that selecting important and reliable features is crucial for multi-modal SOD. Besides, we provide different architectures for comparisons to study the influence, which are only trained with RGB-D data, including:

(a) **Element-wise Sum.** We re-train our model by replacing DCG with an element-wise sum operation at each CMFM.

(b) **Cross-modal Guided Attention Module (CGA).** The cross-modal fusion module is defined in [1]. We also re-train our model by replacing DCG with CGA at each CMFM.

(c) **Dynamic Cross-modal Guided Fusion Module (DCG).** We apply our proposed DCG strategy to dynamically learn cross-modal information from multi-modal features as discussed in Section III-C.

The above configurations are shown in Fig. 4. From Table IX, changing DCG or CGA module into an element-wise sum operation will degenerate saliency result. This indicates that simply adding operation may become easily trapped in complex cross-modal scenes without considering the reliability of complementary modality data. Moreover, the comparisons between CGA module and DCG module show that exploring different sizes of receptive fields is also important for improving the performance in multi-modal SOD tasks. We also visualize saliency maps to intuitively observe the effectiveness

Fig. 11. Visual comparisons of different DCG variants. (a) RGB image. (b) Depth map or thermal image. (c) The saliency results of element-wise sum operation. (d) The saliency results of CGA module. (e) The saliency results of DCG module. (f) GT.

of DCG in Fig. 11. Compared with the element-wise sum operation and CGA module, it is obvious that the DCG fusion strategy achieves better results with less background noises and more complete structures. However, without good understanding on multi-scale cues, CGA may introduce cluttered background or miss the salient region in multiple objects scenes, e.g., the shoe cabinet in the first row or the phone in the last row. These results demonstrate that the design of DCG is crucial for improving saliency performance.

3) **Scalability of MMNet:** We implement our proposed model based on different popular backbones to demonstrate its scalability and effectiveness, including VGG19 [82], ResNet50 [88] and Res2Net50 [83]. In Table X, the “*” denotes that the 1×1 kernel size changes into 3×3 kernel size, which is used to convert the channel of feature maps from backbone to $\{f_{c_i}\}_{i=2}^5$ or $\{f_{r_i}\}_{i=2}^5$. Specifically, all models are trained with RGB-D data for comparison. We can see that the MAE of VGG19 configuration is slightly higher than that of ResNet configuration, while it achieves higher F-measure values. Our proposed method outperforms the reference methods DMRA (VGG19 backbone) on all datasets under different backbones, which shows the scalability and effectiveness of our designed network. Besides, the VGG19 configuration achieves 4.1% and 9.2% improvements in terms of F-measure score on DUT-RGBD and VT5000 datasets compared with the DMRA (VGG19 backbone). Moreover, when implementing ResNet50 like D3Net as the backbone, our method can achieve 41.4% gain in terms of MAE value on LFSD dataset. Moreover, comparing with the DMRA model, our model achieves the

high-accuracy and low-cost multi-modal saliency results under different backbones.

V. CONCLUSION

In this paper, we propose a unified MMNet framework for accurate multi-modal RGB-D and RGB-T salient object detection (SOD). Inspired by the core mechanism of visual color stage doctrine in the human visual system, we propose a cross-modal multi-stage fusion module (CMFM), which correspondingly has a feature response stage and an adversarial combination stage for exploring the effectiveness and complementarity between different modalities. In addition, we propose a bi-directional multi-scale decoder (BMD) to effectively promote the blending of cross-modal features at various scales for better saliency results with complete structure and coherent details. Extensive experiments demonstrate that our proposed method achieves considerable improvements when compared with the other 21 state-of-the-art methods on nine benchmark datasets, which are supported by comprehensive ablation studies. The proposed MMNet shows its robustness and generality on multi-modal RGB-D and RGB-T SOD tasks, making it a benchmark to accelerate the development of this area. Our analyses and benchmark are also expected to inspire novel ideas in related topics, such as RGB-D-T SOD.

REFERENCES

- [1] 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 *the 28th ACM International Conference on Multimedia*, 2020, p. 2436–2444.
- [2] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, 2005.
- [3] K. R. Jeripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1896–1909, 2016.
- [4] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, 2009.
- [5] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *International conference on machine learning*, 2015, pp. 597–606.
- [6] Q. Jiang, F. Shao, W. Lin, K. Gu, G. Jiang, and H. Sun, "Optimizing multistage discriminative dictionaries for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2035–2048, 2017.
- [7] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3089–3098.
- [8] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3085–3094.
- [9] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.
- [10] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *IEEE International Conference on Computer Vision*, 2019, pp. 8779–8788.
- [11] J. Wei, S. Wang, and Q. Huang, "F3net: Fusion, feedback and focus for salient object detection," *arXiv preprint arXiv:1911.11445*, 2019.
- [12] Z. Li, C. Lang, J. Liew, Q. Hou, Y. Li, and J. Feng, "Cross-layer feature pyramid network for salient object detection," *arXiv preprint arXiv:2002.10864*, 2020.
- [13] X. Hu, C.-W. Fu, L. Zhu, T. Wang, and P.-A. Heng, "Sac-net: Spatial attenuation context for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, 2020.
- [14] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, 2020.
- [15] L. Wang, R. Chen, L. Zhu, H. Xie, and X. Li, "Deep sub-region network for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, 2020.
- [16] 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, 2018.
- [17] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *IEEE International Conference on Computer Vision*, 2019, pp. 7254–7263.
- [18] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936.
- [19] 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, 2019.
- [20] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 3051–3060.
- [21] 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, 2019.
- [22] N. Wang and X. Gong, "Adaptive fusion for rgb-d salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.
- [23] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "Pdnet: Prior-model guided depth-enhanced network for salient object detection," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 199–204.
- [24] D.-P. Fan, Z. Lin, J.-X. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, datasets, and large-scale benchmarks," *arXiv preprint arXiv:1907.06781*, 2019.
- [25] 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 *Chinese Conference on Image and Graphics Technologies*. Springer, 2018, pp. 359–369.
- [26] 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 *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019, pp. 141–146.
- [27] 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, 2019.
- [28] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive encoder-decoder network for rgbt salient object detection," *arXiv preprint arXiv:2005.02315*, 2020.
- [29] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbsalient object detection: A large-scale dataset and benchmark," *arXiv preprint arXiv:2007.03262*, 2020.
- [30] G. E. Müller *et al.*, "Darstellung und erklärung der verschiedenen typen der farbenblindheit nebst erörterung der funktion des stabchenapparates sowie des farbensinns der beinen und der fische," 1924.
- [31] G. E. Müller, *Über die Farbenempfindungen*. Joh. Ambr. Barth, 1930.
- [32] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. Siegelbaum, A. J. Hudspeth, and S. Mack, *Principles of neural science*. McGraw-hill New York, 2000.
- [33] K. R. Gegenfurtner, "Cortical mechanisms of colour vision," *Nature Reviews Neuroscience*, vol. 4, no. 7, pp. 563–572, 2003.
- [34] S. G. Solomon and P. Lennie, "The machinery of colour vision," *Nature Reviews Neuroscience*, vol. 8, no. 4, pp. 276–286, 2007.
- [35] B. R. Conway, "Color vision, cones, and color-coding in the cortex," *The neuroscientist*, vol. 15, no. 3, pp. 274–290, 2009.
- [36] G. M. Boynton, "Color vision: how the cortex represents color," *Current Biology*, vol. 12, no. 24, pp. R838–R840, 2002.
- [37] C. Lu and D. H. Fender, "The interaction of color and luminance in stereoscopic vision," *Investigative Ophthalmology & Visual Science*, vol. 11, no. 6, pp. 482–490, 1972.
- [38] M. S. Livingstone and D. H. Hubel, "Psychophysical evidence for separate channels for the perception of form, color, movement, and depth," *Journal of Neuroscience*, vol. 7, no. 11, pp. 3416–3468, 1987.
- [39] N. Kanwisher and E. Wojciulik, "Visual attention: insights from brain imaging," *Nature reviews neuroscience*, vol. 1, no. 2, pp. 91–100, 2000.
- [40] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.
- [41] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of intelligence*. Springer, 1987, pp. 115–141.

- [42] 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, 1998.
- [43] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, 2010.
- [44] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *IEEE conference on Computer Vision and Pattern Recognition*, 2020, pp. 3472–3481.
- [45] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *IEEE conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765.
- [46] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection," *Pattern Recognition*, vol. 86, pp. 376–385, 2019.
- [47] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, 2017.
- [48] 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, 2017.
- [49] J. Guo, T. Ren, and J. Bei, "Salient object detection for rgb-d image via saliency evolution," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.
- [50] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 1115–1119.
- [51] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for rgb-d salient object detection," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 2343–2350.
- [52] 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 Processing Letters*, vol. 23, no. 6, pp. 819–823, 2016.
- [53] 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, 2017.
- [54] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2014.
- [55] L. Zhang, J. Ai, B. Jiang, H. Lu, and X. Li, "Saliency detection via absorbing markov chain with learnt transition probability," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 987–998, 2017.
- [56] Q. Liu, X. Hong, B. Zou, J. Chen, Z. Chen, and G. Zhao, "Hierarchical contour closure-based holistic salient object detection," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4537–4552, 2017.
- [57] Y. Xu, X. Hong, F. Porikli, X. Liu, J. Chen, and G. Zhao, "Saliency integration: An arbitrator model," *IEEE Trans. Multimedia.*, vol. 21, no. 1, pp. 98–113, 2018.
- [58] Y. Xu, D. Xu, X. Hong, W. Ouyang, R. Ji, M. Xu, and G. Zhao, "Structured modeling of joint deep feature and prediction refinement for salient object detection," in *IEEE International Conference on Computer Vision*, 2019, pp. 3789–3798.
- [59] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 454–461.
- [60] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *European conference on computer vision*. Springer, 2014, pp. 92–109.
- [61] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang, "Exploiting global priors for rgb-d saliency detection," in *IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 25–32.
- [62] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3052–3062.
- [63] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8582–8591.
- [64] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "Asif-net: Attention steered interweave fusion network for rgb-d salient object detection," *IEEE Trans. Cybern.*, 2020.
- [65] M. Zhang, S. X. Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, "Asymmetric two-stream architecture for accurate rgb-d saliency detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 374–390.
- [66] D. Liu, Z. Chen *et al.*, "Attentive cross-modal fusion network for rgb-d saliency detection," *IEEE Trans. Multimedia.*, 2020.
- [67] Y. Ma, D. Sun, Q. Meng, Z. Ding, and C. Li, "Learning multiscale deep features and svm regressors for adaptive rgb-t saliency detection," in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1. IEEE, 2017, pp. 389–392.
- [68] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Computational Visual Media*, pp. 1–33, 2021.
- [69] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *European conference on computer vision*. Springer, 2010, pp. 366–379.
- [70] 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, 2014.
- [71] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *IEEE conference on Computer Vision and Pattern Recognition*, 2020, pp. 4003–4012.
- [72] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
- [73] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 421–429.
- [74] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813.
- [75] C. Zhu and G. Li, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3008–3014.
- [76] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *IEEE international conference on computer vision*, 2017, pp. 4548–4557.
- [77] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 733–740.
- [78] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1597–1604.
- [79] 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, 2015.
- [80] D. Fan, C. Gong, Y. Cao, B. Ren, M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *The Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, pp. 698–704.
- [81] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *IEEE conference on computer vision and pattern recognition*, 2014, pp. 248–255.
- [82] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [83] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE TPAMI*, 2020.
- [84] 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 *IEEE conference on Computer Vision and Pattern Recognition*, 2020, pp. 9060–9069.
- [85] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 3203–3212.
- [86] V. Koltun *et al.*, "Efficient inference in fully connected crfs with gaussian edge potentials," *Adv. Neural Inf. Process. Syst.*, vol. 4, 2011.
- [87] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 936–944.
- [88] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Wei Gao (M'18) received the Ph.D. degree in computer science from City University of Hong Kong, Kowloon, Hong Kong, February 2017. From 2012 to 2013, he was a Camera ISP Engineer with the OmniVision Technologies, Shanghai, China. In 2016, he was a Visiting Scholar with University of California, Los Angeles (UCLA), CA, USA. From 2017 to 2019, he was a Postdoctoral Fellow with City University of Hong Kong, Hong Kong, and Research Fellow with Nanyang Technological University, Singapore. Since 2019, he has been an

Assistant Professor at the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China, and also affiliated to Peng Cheng Laboratory, Shenzhen, China. His research interests include image and video coding, image and video restoration, point cloud compression and processing, salient and 3D object detection, visual perception modeling, multimedia and AI accelerators.

Yongsheng Liang received the Ph.D. degree in communication and information systems from the Harbin Institute of Technology, Harbin, China, in 1999. From 2002 to 2005, he was a Research Assistant at the Harbin Institute of Technology, where he has been a Professor with the School of Electronic and Information Engineering, since 2017. His research interests include video coding and transferring and machine learning and applications. He was a recipient of the Wu Wenjun Artificial Intelligence Science and Technology Award for Excellence and

the Second Prize in Science and Technology of Guangdong Province.

Guibiao Liao received the B.E. degree from Tianjin University, Tianjin, China, in 2020. He is currently pursuing the M.E. degree at the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China. His research interests include computer vision and deep learning.

Siwei Ma (M'03-SM'12) received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. He held a post-doctoral position with the University of Southern California at Los Angeles, Los Angeles, CA, USA, from 2005 to 2007. He joined the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing, where he is currently a Professor. He has au-

thored more than 200 technical articles in refereed journals and proceedings in image and video coding, video processing, video streaming, and transmission. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video technology and the Journal of Visual Communication and Image Representation.

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

He is a fellow of IEEE and 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 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 IEEE Circuits and Systems Society from 2016 to 2017 and the Asia-Pacific Signal and Information Processing Association (APSIPA) from 2012 to 2013.

Ge Li received the B.E. degree from the Department of Computer Science and Engineering, Dalian University of Technology, Dalian, China, in 1988, and the M.S. degree from the Institute of Acoustics, Chinese Academy of Science, Beijing, China, in 1991. From 1991 to 1992, he was a Lecturer with the Department of Electrical Engineering, Dalian Maritime University, Dalian. In 1995, he was a Summer Engineer with CSG Research Laboratories, Motorola Inc., Libertyville, IL, USA. He is currently

is a Professor with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China. His general research interests include image/video process and analysis, machine learning, digital communications, and signal processing.