Improving RGB-D Salient Object Detection via Modality-Aware Decoder

Mengke Song, Wenfeng Song, Guowei Yang, and Chenglizhao Chen, *Member, IEEE*

***Abstract*— Most existing RGB-D salient object detection (SOD) methods are primarily focusing on cross-modal and cross-level saliency fusion, which has been proved to be efficient and effective. However, these methods still have a critical limitation, *i.e.*, their fusion patterns – typically the combination of selective characteristics and its variations, are too highly dependent on the network's non-linear adaptability. In such methods, the balances between RGB and D (Depth) are formulated individually considering the intermediate feature slices, but the relation at the modality level may not be learned properly. The optimal RGB-D combinations differ depending on the RGB-D scenarios, and the exact complementary status is frequently determined by multiple modality-level factors, such as D quality, the complexity of the RGB scene, and degree of harmony between them. Therefore, given the existing approaches, it may be difficult for them to achieve further performance breakthroughs, as their methodologies belong to some methods that are somewhat less modality sensitive. To conquer this problem, this paper presents the Modality-aware Decoder (MaD). The critical technical innovations include a series of feature embedding, modality reasoning, and feature back-projecting and collecting strategies, all of which upgrade the widely-used multi-scale and multi-level decoding process to be modality-aware. Our MaD achieves competitive performance over other state-of-the-art (SOTA) models without using any fancy tricks in the decoder's design. Codes and results will be publicly available at https://github.com/MengkeSong/MaD.**

***Index Terms*— RGB-D salient object detection, modality-aware fusion, deep learning.**

I. Introduction and Motivation

SALIENT Object Detection (SOD) aims at well-segmenting the most eye-attracting objects in a given image scene, and this topic has been extensively studied for over 20 years. Thanks to the rapid development of deep learning technology, it has been widely used in various computer vision-related downstream applications, such as image retrieval [1], [2], [3], [4], image translation [5], [6], object/face detection [7], [8], [9], [10], [11], segmentation [12], [13], compression [14], [15], and even video tracking [16], [17]. Albeit making significant progress, we have observed that new performance improvement achieved by the recent works [18], [19], [20], [21], [22] shrinks significantly, indicating the solely RGB image-based SODs have reached a performance bottleneck.

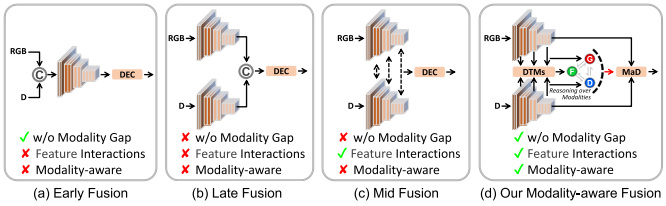


Fig. 1. Pictorial demonstrations of the existing RGB-D fusion schemes. We use ✓ and ✗ to denote advantages and disadvantages.

Different from the single RGB SODs [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], RGB-D SODs [34], [35], [36], [37], [38] have become a scorching research topic recently since depth-sensing cameras are more accessible than ever before, *e.g.*, even for a smart mobile phone, depth-sensing cameras have been widely equipped [39]. It is both an opportunity and a challenge since the additional D information enhances the potential to achieve a further SOD performance improvement, yet designing an appropriate fusion logic is also not an easy task.

In general, the fusion methods adopted by **s**tate-**o**f-**t**he-**a**rt (SOTA) RGB-D SODs can be categorized into three groups: 1) early fusion [41], [42], 2) late fusion [43], [44], and 3) mid fusion [45], [46], [47], [48], [49], [50], as illustrated in Fig. 1. Though early fusion and late fusion have their advantages, they usually perform poorly due to the absence of feature interaction. Thus the current mainstream SOTA models have widely adopted the mid-fusion. Nevertheless, such a widely-used fusion scheme has a critical limitation, *i.e.*, the fusion process is 'modality-unaware[1]' or

Manuscript received 17 February 2022; revised 16 July 2022 and 1 September 2022; accepted 1 September 2022. Date of publication 16 September 2022; date of current version 22 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62172246 and Grant 62172229, in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems under Grant VRLAB2021A05, and in part by the Youth Innovation and Technology Support Plan of Colleges and Universities in Shandong Province under Grant 2021KJ062. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mai Xu. *(Corresponding author: Chenglizhao Chen.)*

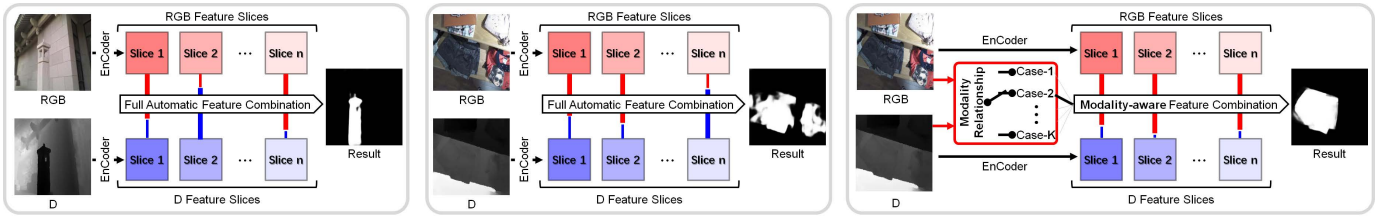
Mengke Song and Chenglizhao Chen are with the College of Computer Science and Technology and the Qingdao Institute of Software, China University of Petroleum (East China), Qingdao 266580, China (e-mail: cclz123@163.com).

Wenfeng Song is with the Computer School, Beijing Information Science and Technology University, Beijing 100192, China.

Guowei Yang is with the School of Electronic Information, Qingdao University, Qingdao 266071, China.

Digital Object Identifier 10.1109/TIP.2022.3205747

[1]Here we use the term 'modality-unaware' to highlight that the modality relationship was inappropriately learned by the existing SOTA works. We are also fully aware that the current methods can somehow learn the modality relationship, but not as appropriate as our approach.



(a) **Modality-unaware** Mid Fusion in **Common Scene** (b) **Modality-unaware** Mid Fusion in **Uncommon Scene** (c) Our **Modality-aware** Approach in **Uncommon Scene**

Fig. 2. Real instances to illustrate the advantage of our modality-aware fusion against the widely-used modality-unaware mid fusion. In view of the conventional mid fusion (DMRA [40]), the RGB-D combination rules learned in common senses (a) might not be suitable for uncommon scenes (b), and such a problem can be solved only if the modality relationship has been considered in advance, *e.g.*, (c).

'modality-near-aware'. An optimal complementary fusion status between RGB and D is often determined by multiple modality-related factors, *e.g.*, D quality, the complexity of the RGB scene, and harmony degree between RGB and D; all such factors are objective, and they might not be perceived by the current SOTA models, because their fusion logics simply focus on learning non-linear mapping over intermediate features derived by off-the-shelf encoders (*e.g.*, ResNet [51]).

To facilitate a better clarification, Fig. 2 provides some pictorial instances. From the perspective of mid-fusion, the problem is that those RGB-D combination rules are learned from ordinary RGB senses (a). And the majority of training instances in real works usually adapt poorly in facing uncommon scenes (RGB saliency conflicts with D saliency), *e.g.*, (b). The main reasons are twofold: 1) those uncommon scenes are usually the minorities, leading to biased training; 2) more importantly, such mid fusion often formulates its RGB-D combination rules by considering intermediate feature slices individually — a very local methodology, leading to the training task to be very complex and, eventually, making the biased learning situation worse.

As shown in the first two rows of Fig 3, a typical case whose depth shall play a dominant role, the modality relationship in such case cannot be well learned by the existing methods. Thus the salient object has been detected inaccurately. Also, when the RGB takes hold in the detection shown in the last two rows, the existing SOTA methods still perform not very well, which are inferior to our modality-aware method in such cases. The main reason is that these methods do not appropriately learn the complemental relationship between RGB and depth; they merely integrate the two-modality feature slices in a local manner. The harmony degree between RGB and depth determines the degree of complementarity, and the harmony degree is what we call '**modality-aware**'.

In contrast to the existing works, the major highlight of our approach, a novel **M**odality-**a**ware **D**ecoder (MaD) demonstrated in Fig. 2 (c), is to reason between different modalities. The learned modality relationships will later be used to guide the subsequent dense feature collection, making the decoding process focus more on the modality relationship and thus perform well in those uncommon scenes.

More theoretically speaking, our approach divides all potential RGB-D combinations into coarse-granularity 'boxes/cases' in advance (Fig. 2-c), and thus the subsequent fine-granularity slice-level combinations can be learned correctly. The proposed MaD mainly consists of two parts, *i.e.*, 1) a feasible way to decouple and represent RGB-D images into a modality-related feature space, which solids the basis of our MaD, and will be detailed in Sec. III-B, and 2) a series of feature embedding and back-projecting strategies are devised to convert the SOTA decoding process to be modality-aware, and this part will be detailed in the rest of Sec. III.

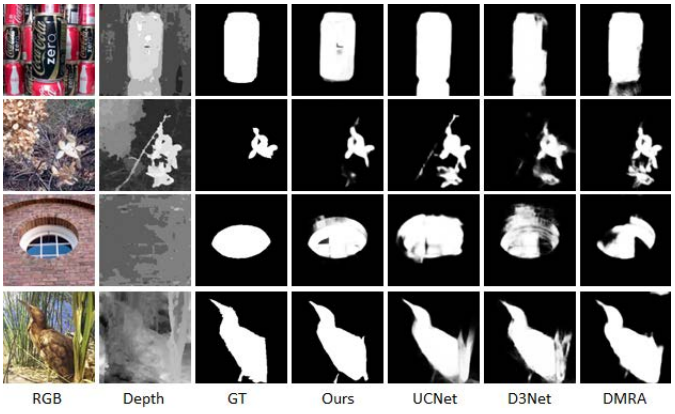


Fig. 3. Visualized comparison among our method and some modality-unaware methods (UCNet [52], D3Net [53] and DMRA [40]). The first two lines represent complex RGB scene with clear depth, while the last two lines denote clear RGB scene with low-quality depth.

The key contributions of this paper can be summarized in the following three aspects:

- As one of the first attempts, this paper has raised attention regarding the importance of modality relationships in performing RGB-D SOD. In addition, some in-depth discussions and explanations of this issue have been conducted.
- This paper has devised a feasible solution to perform modality-level reasoning to achieve modality-aware RGB-D fusion. A series of functional modules are presented to facilitate the use of modality relationships to guide multi-scale/level feature collection.
- We have conducted extensive experiments to verify the effectiveness and superiority of our approach (*i.e.*, MaD). Both codes and results are now publicly available, potentially benefiting our research community in the near future.

II. Related Work

A. RGB-D Salient Object Detection

Traditional methods mainly rely on hand-craft features [54], [55], [56], [57], [58], using a large amount of saliency prior information for image saliency detection, such as contrast prior, image background prior, target prior, and so on. Zhu and Li [54] proposed a multi-layer back-propagation saliency detection algorithm based on depth mining to exploit depth cues from three different layers of images. Zhu *et al.* [55] used a center-dark channel prior by generating a center-dark channel map based on a center saliency prior and a dark channel prior. They then fused the initial saliency map with the center-dark channel map to generate the final saliency map. This algorithm is straightforward and can also be applied to small object detection. However, these methods seem simple but disregard the differences between the RGB and depth modalities and thus might not achieve reliable results.

As deep learning comes into view, many CNNs-based methods [50], [59], [60], [61], [62], [63], [64], [65], [66], [67] dominate this field. Among them, fusion-based methods [41], [42], [43], [44], [53], [68], [69], [70], [71], [72], [73] have devoted significantly to RGB-D saliency detection and have achieved appealing performance. As is shown in Fig. 1, for the first category, input fusion [41], [42], [69] refers to directly serializing the RGB image and depth map to form a four-channel RGB-D input. Especially, Song *et al.* [41] performed multi-scale pre-segmentation on the RGB-D pairs and proposed the multi-scale discriminative saliency fusion to generate the final saliency map. In terms of late fusion [43], [44], Guo *et al.* [43] iteratively propagated the initial saliency map, which is produced by multiplication, to generate the final saliency map. Considering the quality of the depth map, Cong *et al.* [44] proposed a measure to evaluate the reliability of the depth map and used it to combine the two predictions. Meanwhile, for middle fusion [53], [70], [71], [72], [73] which adopts a two-stream structure to convert cross-modal features and fuse cross-level features, Fan *et al.* [53] used a gate mechanism to filter out the low-quality depth maps explicitly. Li *et al.* [72] proposed to fuse high-level RGB and depth features interactively and adaptively, discriminate the cross-modal features from different sources, and enhance RGB features with depth features at each level. Chen *et al.* [73] integrated a depth quality-aware subnetwork into the classic bi-stream structure, assigning the weight of the depth feature before conducting the fusion. Though these methods gain tremendous achievements in performance, they still encounter problems of incomplete feature fusion.

Most existing fusion-based RGB-D SOD methods mainly adopt depth clues as supplements to assist the RGB branch since RGB and depth might play a disparate role in different scenes, *e.g.*, clear depth and complex RGB *vs.* clear RGB and fuzzy depth. Thus, apart from devising elaborate fusion approaches to merge them, it's also necessary to take the two modalities individually, that is, 'modality-aware'. Zhai *et al.* [70] have presented a bifurcated backbone strategy network. In this work, the proposed network follows a bifurcated backbone strategy to recombine multi-level features into teacher and student features and integrate RGB and depth modalities in a complementary manner. Specifically, the authors have proposed a depth-enhanced module to excavate informative depth cues from the channel and spatial views, achieving significant performance gain. Also, Wang *et al.* [46] have proposed a simple yet effective method to learn discriminative and modality-specific features for RGB and depth and explicitly extract useful, consistent information from them. Also, regarding the cross-modality consistency, the authors have calculated the correlations of every pixel pair from RGB (RGB correlation) or depth inputs (depth correlation). Recently, Zhou *et al.* [74] have devised a novel specificity-preserving network that explores shared information and modality-specific properties. Given the methodology innovation, the authors have utilized two modality-specific networks and a shared learning network to generate individual and shared saliency maps, yielding better segmentation results by a cross-enhanced integration and a multi-modal feature aggregation operation.

However, these modality-unaware or modality-near-aware methods focus more on the modality-specific properties regarding cross-level or cross-modality fusion in the encoding phase, which might neglect that the decoding phase is required to be modality-aware. Based on this insight, we propose a modality-aware decoder to complete the whole training phase of the encoder-decoder.

B. Attention Mechanism

Attention mechanisms [63], [75], [76], [77], [78] were introduced into computer vision to imitate the aspect of the human visual system. Such an attention mechanism can be regarded as a dynamic weight adjustment process based on features of the input image.

Kuen *et al.* [79] firstly introduced an attention mechanism to salient object detection tasks. They used spatial transformer and recurrent network units to iteratively attend to selected image sub-regions to perform saliency refinement progressively and learn context-aware features from past iterations to enhance saliency refinement in future iterations. Zhai *et al.* [70] proposed a depth-enhanced module consisting of sequential channel attention and spatial attention to excavating informative parts of depth cues from the channel and spatial views. Liu *et al.* [80] integrated the self-attention and each other's attention to propagate long-range contextual dependencies to incorporate multi-modal information to learn attention and propagate contexts more accurately and selection attention to weight the newly added attention term. Li *et al.* [81] proposed an alternate interaction unit composed of several channel attention and spatial attention operations to filter distracters' in-depth features. Then the purified depth features were exploited to enhance RGB features in turn.

Meanwhile, as is different from [82] employing depth information as an error-weighted map to correct the segmentation process, we utilize integrated information as a weighting vector to guide the following fusion processing procedures. We can learn the modal-level relationships between the two modalities by employing modality-aware dynamic fusion.

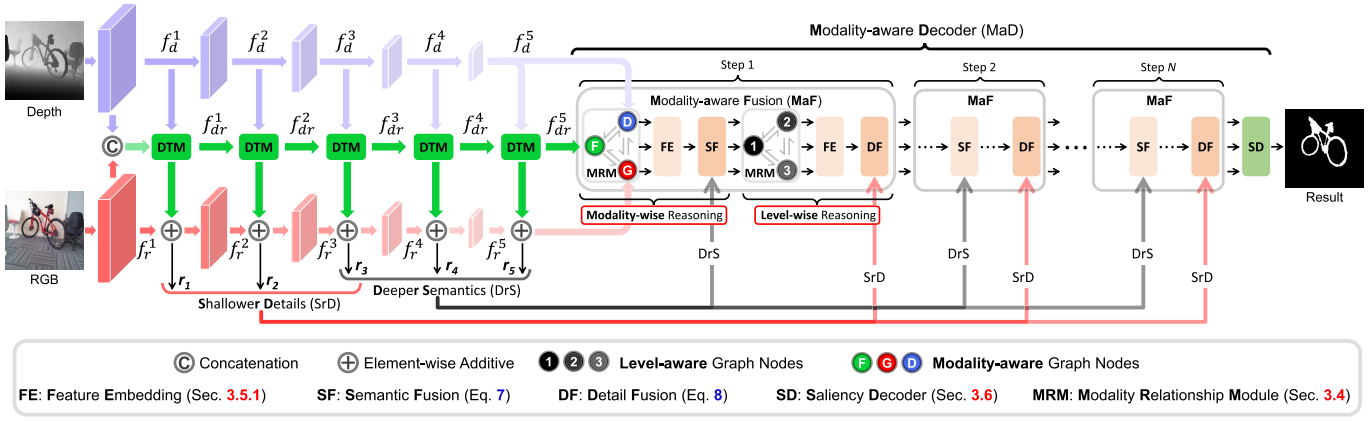


Fig. 4. The overall method pipeline, where the **m**odality-**a**ware **d**ecoder (MaD) is the major highlight of this paper. The proposed MaD targets at learning the relationship between different modalities, which can promote the later multi-level RGB-D fusion process by letting the network to be aware whether a fused RGB-D feature channel is helpful to the current **s**alient **o**bject **d**etection (SOD) task.

C. Graph Neural Network

A GNN can be viewed as a message-passing algorithm, where representations for nodes are iteratively computed conditioned on their neighboring nodes through a differentiable aggregation function. Xia and Gao [83] proposed a dense graph convolution to enhance the local context information of joints, and spatial and temporal attention modules are used to adapt the intermediate feature maps. Xu *et al.* [84] constructed a super-pixel level spatiotemporal graph among multiple frame-pairs and imported graph data into the devised multi-stream attention-aware GCN. Luo *et al.* [85] proposed to distill and reason the mutual benefits between these RGB data and depth data sources through cascade graphs. Jiang *et al.* [86] propagated information across multiple graphs and obtained a consistent representation and learning by integrating the information of multiple graphs simultaneously.

Unlike all the above methods, our MaD introduces a feature weighting strategy to guide feature fusion and yield purified all-sided features given cross-modal and cross-level.

III. Proposed Method

A. Method Overview

Fig. 4 shows the overall pipeline of our approach, which mainly includes two parts, illustrated from left to right: 1) a mid-fusion encoder supported by **D**epth **T**ransfer **M**odule (DTM), and 2) a novel **M**odality-**a**ware **D**ecoder (MaD) consisting of N sequential **M**odality-**a**ware **F**usions (MaF). The part 1) takes two different modalities (*i.e.*, RGB and D) as input to perform mid-level saliency dense fusion, where the DTM narrows the feature gap between different modalities and serves as an intermediate modality. Technical details regarding part 1), which embeds RGB and D to a uniform feature space, will be provided in Sec. III-B. The part 2), also the highlight of this paper, targets a learning modality-aware relationship to guide the multi-scale and multi-level feature collection, which, compared with the conventional encoders (*e.g.*, [72], [73]), can achieve better complementary RGB-D fusion. And this part will be detailed in Sec. III-C.

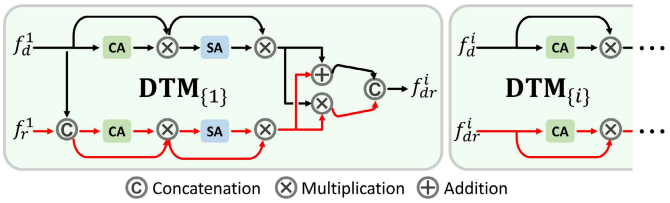


Fig. 5. Detailed architecture of **d**epth **t**ransfer **m**odule (DTM). CA: channel attention; SA: spatial attention.

B. Mid-Level RGB-D Fusion via DTM

In general, when performing RGB-D saliency fusion, a vital issue is that low-quality D — a widespread phenomenon in RGB-D images, could bring noises or erroneous interruptions to the fusion process resulting in poor **s**alient **o**bject **d**etection (SOD) results. Therefore, instead of performing complete dense RGB-D fusion at a time, the encoder shall use D to complement RGB conservatively, *i.e.*, those D regions which contradict their RGB counterparts should be temporally omitted. As such, we propose the **D**epth **T**ransfer **M**odule (DTM), which consists of multiple multiplicative-based fusion operations to compress those contradicted RGB-D regions. The proposed DTM targets two objectives: 1) mining high-quality and practical D, and 2) using them to promote their RGB counterparts; the corresponding technical details have been shown in Fig. 5.

We use f_r^i/f_d^i to represent features derived from the i-th layer of RGB/D encoder. DTM mainly consists of two streams, *i.e.*, the D stream (black arrows) and the fused RGB-D stream (red arrows). Each stream performs multiplicative operation-based feature enhancing in advance, *e.g.*, residual operations, spatial- and channel-wise attentions. Both streams will be combined via parallel addition and multiplicative operations. Notice that the multiplicative operation in the last parallel fusion stage could effectively compress inconsistencies between RGB and D, which can be very effective in mining high-quality and practical D. Both **s**patial-wise **a**ttention (SA) and **c**hannel-wise **a**ttention (CA) adopted in DTM follow the

typical ways, which can be respectively represented as:

$$\mathrm{CA}(f) = moc\Big(f, \mathrm{MLP}\Big[\mathrm{GMP}_c(f)\Big]\Big), \tag{1}$$

where f denotes the input feature, GMP_c is the global max pooling operation over the input feature slice, MLP stands for a two-layer perception, and $moc(\cdot,\cdot)$ performs channel-wise multiplication between its input;

$$\mathrm{SA}(f) = ewm\Big(f, \mathrm{Conv3}\Big[\mathrm{GMP}_s(f)\Big]\Big), \tag{2}$$

GMP_s is the pixel-wise global max-pooling over the entire input feature tensor, Conv3 is a 3×3 convolution, and $ewm(\cdot,\cdot)$ performs element-wise multiplication between input.

Notice that, as shown in Fig. 4, there are a total of 5 DTMs adopted in the mid-level fusion process, where each DTM correlates to an individual encoder level. Clearly, such cascade DTMs have two different inputs, where Fig. 5 illustrates the difference, *i.e.*, the input of $\mathrm{DTM}_{\{1\}}$ includes both f_r^i and f_d^i, yet the input of $\mathrm{DTM}_{\{i\}}$ consists of f_d^i and f_{dr}^i, where f_{dr}^i is the output of $\mathrm{DTM}_{\{i-1\}}$. Hence, the output of DTM can be represented as:

$$f_{dr}^i = \begin{cases} \mathrm{DTM}_{\{i\}}\left(f_d^1, f_r^1\right) & if\ i = 1 \\ \mathrm{DTM}_{\{i\}}\left(f_d^i, f_{dr}^{i-1}\right) & if\ i = \{2,3,4,5\}. \end{cases} \tag{3}$$

Here we take the 3rd DTM for instance, and its inside dataflow can be briefed as:

$$\begin{aligned} \mathrm{DTM}_{\{3\}}(f_d^3, f_{dr}^2) &= \mathrm{Conv1}\Big[\mathcal{C}\Big(F_d^3 + F_{dr}^2, F_d^3 \otimes F_{dr}^2\Big)\Big], \\ F_{dr}^2 &= \mathrm{SA}\big(\mathrm{CA}(f_{dr}^2) \times f_{dr}^2\big) \times \big(\mathrm{CA}(f_{dr}^2) \times f_{dr}^2\big), \\ F_d^3 &= \mathrm{SA}\big(\mathrm{CA}(f_d^3) \times f_d^3\big) \times \big(\mathrm{CA}(f_d^3) \times f_d^3\big), \end{aligned} \tag{4}$$

where f_{dr}^2 is the output of $\mathrm{DTM}_{\{2\}}$, and Conv1 stands for a 1×1 convolution. Next, each output of the cascade DTMs, *i.e.*, the high-quality f_{dr}^i, will be used to promote features embedded in the RGB decoder via addition operation.

By performing the abovementioned procedure, we have decoupled the RGB-D saliency fusion process into three individual parts, *i.e.*, f_d^i, f_{dr}^i, and f_r^i, which respectively stand for 1) raw D features, 2) consistency degree between RGB and D, and 3) RGB features conservatively enhanced by D. Though the RGB features have been conservatively enhanced, this fusion process is still modality-unaware, and some valuable D currently conflicting with RGB are still embedded in f_d^i and f_{dr}^i. Thus, in the following subsection, we will introduce a feasible way to model the relationship between different modalities, aiming to achieve modality-aware RGB-D fusion, which is critical for obtaining better complementary fusion status between RGB and D.

C. *Modality-aware Decoder (MaD)*

Most of the existing cross-modal fusion methods [87], [88], [89], [90] follow the conventional selective fusion methodology, where their fusion processes assume that RGB and D are equally helpful towards the SOD task, and the balances between RGB and D are online learned by their network under the guidance of the given learning objective. Such a learning process is very 'slack' and modality-unaware. Yet, the fact is that the optimal RGB-D complementary status is usually determined by multiple modality-related aspects, *e.g.*, RGB scene complexity, and D quality. Consequently, achieving an optimal balance between RGB and D might be tough without considering such objective aspects. Thus, we present the MaD, composed of N **M**odality-**a**ware **F**usion modules (MaF), which can collect multi-scale and multi-level features in a modality-aware way. As is shown in the right part of Fig. 4, MaF mainly consists of two parts: **Modality-wise** Reasoning and **Level-wise** Reasoning. The primary objective of these two parts is to stay modality-aware when performing multi-level feature collection. Both of Modality-wise Reasoning and Level-wise Reasoning include three parts, *i.e.*, **M**odality **R**elationship **M**odule (MRM), **F**eature **E**mbedding (FE), **S**emantic **F**usion (SF)/**D**etail **F**usion (DF).

The relationships between different modalities can be automatically formulated using MRM by performing 1D convolution over a modality graph that contains three nodes, *i.e.*, G, D, and F, which respectively represent RGB modality, D modality, and an additional in-between modality, and the technical details of this part will be provided in Sec. III-C.1.

The rationale of Level-wise Reasoning is quite similar to that of Modality-wise Reasoning. The significant difference between them is that Modality-wise Reasoning directly takes the outputs of the encoder as input (*i.e.*, f_d^5, f_{dr}^5, and f_r^5). In contrast, the input of Level-wise Reasoning is the output of Modality-wise Reasoning, where the learned modality relationship guides the deep semantic feature fusion via MaF. Specifically, MRM learns relationships between different modalities, and FE projects the previously learned modality relationship from the graph interaction space back to the spatial coordinate space. Thus the embedded modality relationship can be used to weight multi-level features (*i.e.*, r_i). SF and DF take both the output of FE and r_i as input to achieve modality-aware multi-scale/level feature collection.

*1) **M**odality **R**elationship **M**odule (MRM):* To dynamically learn the relationships between different modalities, we propose the MRM. As seen in Fig. 4, MRM performs graph convolution over three nodes, *i.e.*, G, D, and F, where the in-between modality F, the output of DTM, is very slack, and it can well reflect the harmony degree between RGB and D — could be very useful to determine their complementary status. By formulating such a relationship and using it to guide the subsequential multi-level decoding process, we can achieve better complementary status between RGB and D regarding the given SOD task.

In our research community, multiple graph convolutions exist to learn the relationships between graph nodes, *e.g.*, the classic GloRe [91]. In our MRM, we regard the output of the last encoder layers as the graph nodes, *i.e.*, f_d^5, f_{dr}^5, and f_r^5, because such outputs are all semantics that are very helpful in perceiving the relationship between different modalities. Meanwhile, since f_d^5, f_{dr}^5, and f_r^5 have already been projected into a uniform feature space via the encoder, their relationship

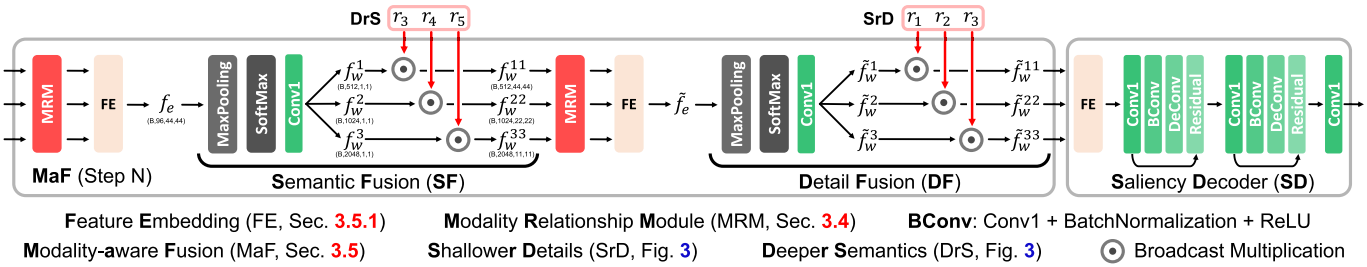


Fig. 6. Detailed architecture of SF and DF. Notice that a more fancy SD could bring some additional performance gain.

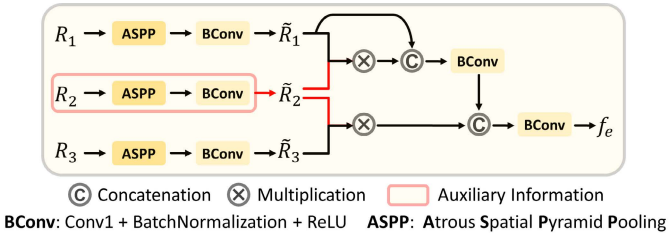


Fig. 7. Architecture of Feature Embedding (FE).

can be simply learned by using two sequential 1D convolutions, and MRM can be detailed as:

$$\{R_1, R_2, R_3\} \leftarrow \mathrm{MRM}(f_d^5, f_{dr}^5, f_r^5) = \mathrm{Conv1D}\Big[\mathrm{Conv1D}\Big[\mathcal{C}\big(f_d^5, f_{dr}^5, f_r^5\big)\Big]\Big], \quad (5)$$

where $\{R_1, R_2, R_3\}$ denote the outputs of MRM, which respectively are the embedded modality-aware features; $\mathcal{C}$ is the typical feature concatenation operation, and Conv1D is a 1D convolution. MRM projects its input into an interactive space for graph nodes, and the relationships between different modalities have been implicitly embedded into it. Notice that a more fancy graph convolution (*e.g.*, [91], [92], [93]) could further promote such features' relationship embedding, yet, to focus on the main topic of this paper, we shall omit other choices.

After obtaining $\{R_1, R_2, R_3\}$, we use them as the dynamic fusion weights to drive the **m**odality-**a**ware **f**usion (MaF), where $\{R_1, R_2, R_3\}$ are used to guide multi-level feature collection towards the side output of the RGB encoder, *i.e.*, r_i demonstrated in Fig. 4. Both method rationale and technical details of MaF will be provided in the next subsection.

2) Feature Embedding (FE): The primary target of FE is to perform feature projection, *i.e.*, $\{R_1, R_2, R_3\} \rightarrow f_e$, where f_e is the learned modality relationship projected back in coordinate space, and thus f_e can be used to weight the subsequential feature collection. Intuitively, all $\{R_1, R_2, R_3\}$ can be simply combined via feature concatenation. However, since $\{R_1, R_2, R_3\}$ respectively correlate to different modalities with varying importance towards the SOD task, we shall treat R_2 as an auxiliary part to complement R_1 and R_3 in a cascade way.

We have demonstrated the technical details of FE in Fig. 7, where the cascaded fusion follows a top-down direction. We also apply **a**trous **s**patial **p**yramid **p**ooling (ASPP) to each stream, which enables multi-scale feature representation in a very efficient way. The dataflow of FE can be formulated as:

$$f_e = \mathrm{BConv}\Big[\mathcal{C}\Big(\mathrm{BConv}\Big[\mathcal{C}(\tilde{R}_1, \tilde{R}_1 \times \tilde{R}_2)\Big], \tilde{R}_2 \times \tilde{R}_3\Big)\Big],$$
$$\tilde{R}_i = \mathrm{BConv}\Big[\mathrm{ASPP}\big(R_i\big)\Big], \quad i \in \{1, 2, 3\}, \quad (6)$$

where $\mathcal{C}$ denotes the feature concatenation operation, BConv sequentially performs 3×3 convolution, batch normalization, and ReLU operations, f_e is the output of FE sharing an identical size with R_i.

This way, the previously learned relationships between different modalities have been embedded in f_e. Thus we can use it to guide multi-level feature collection, *i.e.*, SF, and DF, which will be detailed in the next subsections.

3) Semantic Fusion (SF) and Detail Fusion (DF): We have demonstrated the technical details of both SF and DF in Fig. 6. SF takes f_e, the embedded modality relationship, as its input, and a sequential of MaxPooling, SoftMax, and Conv1 operations are applied to convert f_e into useable attention tensors with desired sizes, *i.e.*, f_w^i.

Meanwhile, the side outputs of the RGB stream illustrated in Fig. 4, *i.e.*, r_i, have included partial D information. Thus, given a conventional saliency decoder (*e.g.*, CPD [94]), one can collect them in a multi-level manner to formulate final saliency maps. The limitation of such a modality-unaware decoder has been mentioned multiple times before. In sharp contrast to the conventional decoders, the f_e is additionally available in our approach, which can make the decoding process modality-aware.

In view of the SOD task, correctly localizing salient objects is usually more critical than achieving boundary salient object segmentation, and features derived in deeper layers are more helpful than those in shallower layers towards this aspect. Therefore, as shown in SF of Fig. 6, we respectively apply f_w^i as attentions on deeper semantic features r_3, r_4, r_5, where modality-aware high-level semantic features (*i.e.*, $f_w^{11}, f_w^{22}, f_w^{33}$) can be obtained accordingly. This process can be detailed as the following equation, in which we take f_w^{11} for instance.

$$f_w^{11} = r_3 \odot f_w^1, \quad f_w^1 = \mathrm{Conv1}\Big[\mathrm{SoftMax}\big(\mathrm{MaP}(f_e)\big)\Big], \quad (7)$$

where $\odot$ denotes the broadcast multiplication, Conv1 is a 1×1 convolution which resizes its input to any wanted size; MaP is the typical max-pooling operation, and softmax operation ensures that f_w^1 can be used as an attention tensor.

TABLE I

QUANTITATIVE EVALUATION OF MAJOR COMPONENTS USED IN OUR APPROACH, WHERE ✓: 'USING', ✗: 'WITHOUT USING'. THE FULL DESCRIPTIONS REGARDING MARKS FROM ① TO ④ CAN BE FOUND IN SEC. IV-E

		Major Components										Datasets and Quantitative Metrics															
		Backbone		MRM		Fusion			Hyper Settings			NJUD				NLPR				LFSD				SSD			
		B	DTM	MR	LR	FE	SF/DF	LwR	IN	En3	En5	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓
①	1	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	.880	.820	.876	.063	.891	.883	.877	.035	.840	.847	.849	.079	.834	.778	.852	.068
	2	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	.902	.851	.888	.049	.911	.885	.891	.029	.855	.866	.869	.066	.847	.804	.879	.060
②	3	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓	.919	.901	.912	.040	.930	.899	.940	.024	.880	.875	.899	.053	.870	.848	.905	.046
	4	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓	.918	.899	.911	.041	.931	.898	.938	.024	.879	.876	.894	.054	.865	.843	.902	.048
	5	✓	✓	✗	✗	✓	✓	✓	✓	✗	✓	.908	.875	.909	.047	.923	.884	.925	.029	.876	.869	.859	.058	.862	.841	.900	.052
③	6	✓	✓	✓	✓	✗	✓	✓	✓	✗	✓	.909	.875	.913	.045	.915	.888	.934	.027	.861	.870	.886	.062	.860	.839	.897	.052
	7	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	.915	.890	.919	.041	.922	.891	.939	.024	.870	.879	.892	.053	.864	.843	.901	.050
	8	✓	✓	✓	✗	✓	✓	✗	✓	✗	✓	.917	.899	.924	.041	.924	.890	.942	.026	.875	.880	.900	.055	.868	.848	.903	.047
④	9	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	.916	.898	.925	.040	.926	.893	.947	.027	.879	.874	.898	.055	.861	.844	.898	.049
	10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	.919	.900	.928	.039	.928	.897	.950	.025	.881	.875	.899	.053	.865	.847	.902	.047
	11	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	**.921**	**.903**	**.930**	**.037**	**.933**	**.901**	**.955**	**.022**	**.884**	**.877**	**.901**	**.051**	**.872**	**.850**	**.907**	**.045**

B: baseline model
DTM: Depth Transfer Module (Fig. 4)
MR: using MRM in Modality-wise Reasoning
LR: using MRM in Level-wise Reasoning
En3/En5: MRM takes the 3th/5th feature of the encoder as input
FE: Feature Embedding (Fig. 5)
SF/DF: Semantic Fusion/Detail Fusion (Fig. 6)
LwR: using Level-wise Reasoning (Fig. 3)
IN: sequentially iterate MaD N times

To achieve a multi-level feature collection, the rest of the encoder's side outputs (*i.e.*, r_1 and r_2) should also be included. Intuitively, both r_1 and r_2 can be simply combined with the output of SF, *e.g.*, f_w^{11}. However, since $f_w^{11}, f_w^{22}, f_w^{33}$ are pretty redundant and their primary values are to localize salient objects, it could lead to incomplete SOD results if such redundant feature responses are not removed. As such, we further explore the relationship between $f_w^{11}, f_w^{22}, f_w^{33}$, an additional Level-wise Reasoning, which can filter such redundant information.

As illustrated between SF and DF in Fig. 6, an additional MRM has been applied over the output of SF, where all $f_w^{11}, f_w^{22}, f_w^{33}$ have been projected back to graph interaction space again and latterly projected back to the coordinate space via FE, *i.e.*, $\tilde{f}_e$. Our DF can be detailed as the following equation, where we take the top line for instance:

$$\begin{aligned} \tilde{f}_w^{11} &= r1 \odot \tilde{f}_w^{1}, \quad \tilde{f}_w^{1} = \mathrm{Conv1}\Big[\mathrm{SoftMax}\big(\mathrm{MaP}(\tilde{f}_e)\big)\Big], \\ \tilde{f}_e &= \mathrm{FE}\Big(\mathrm{MRM}\big(f_w^{11}, f_w^{22}, f_w^{33}\big)\Big), \end{aligned} \tag{8}$$

where $\odot$ denotes the broadcast multiplication, the details of FE and MRM can be found respectively in Sec. III-C.2 and Eq. 5. We can achieve modality-aware and multi-level feature collection using both SF and DF, and the MaF will be repeated N times. As seen in Fig. 6, the output of the N-th MaF will be directly fed to a **s**aliency **e**ncoder (SD) to obtain final saliency prediction. The technical details of SD will be given in the next subsection.

D. *Saliency Decoder (SD)*

After repeating MaF N times, we can get three dynamic fusion features, *i.e.*, $\tilde{f}_w^{11}, \tilde{f}_w^{22}, \tilde{f}_w^{33}$, which have already been embedded in cross-scale and cross-level modality relationships. To produce the final saliency map, we propose the SD. Instead of collecting by up-sampling multi-scale and multi-level features as the typical encoder does, the proposed SD shall only focus on: simultaneously combining $\tilde{f}_w^{11}, \tilde{f}_w^{22}, \tilde{f}_w^{33}$ and performing up-sampling. The architectural detail of SD can be seen in the right part of Fig. 6.

IV. EXPERIMENTS

A. *Datasets*

We evaluate the effectiveness of our model on six widely used public benchmark datasets, *i.e.*, NJUD [95], NLPR [96], SIP [53], STEREO [97], LFSD [98], and SSD [99]. NJUD [95] includes 2,003 stereo image pairs with various resolutions. Among these image pairs, 1,400 are used as the training set, 100 as the validation set, and the remaining as the testing set. NLPR [96] consists of 1,000 images from 11 types of indoor and outdoor scenes. Among them, 650 are used as the training set, 50 as the validation set, and the remaining 300 as the testing set. SIP [53] consists of 1,000 high-resolution images that cover diverse real-world scenes from various viewpoints, poses, occlusions, illuminations, and backgrounds. STEREO [97] has 797 stereoscopic images. These images are mainly collected from the Internet and 3D movies. Depth images are generated by leveraging an optical method. LFSD [98] is a relatively small dataset for testing, which contains 100 images with depth information captured via a Lytro light field camera and manually labeled ground truths. The resolutions of these images are relatively small. SSD [99] has 80 images picked up from stereo movies.

B. *Evaluation Matrices*

Three metrics are adopted for quantitative evaluation, including **S**-measure (Sm) [100], **F**-measure (Fm) [101], E-measure (Em) [102], and mean absolute error (MAE). Specifically, S-measure is utilized to solve the problem of structural measurement from the perspective of region-aware and object-aware. F-measure offers a unified solution to evaluating non-binary and binary maps. E-measure combines local pixel values with the image-level mean value to jointly evaluate the similarity between the prediction and the ground truth. The MAE denotes the average pixel-wise difference between saliency maps and the ground truth.

C. *Implementation Details*

We implemented MaD by PyTorch with an NVIDIA GeForce 2080 GPU. Following [35], [64], the proposed MaD

TABLE II

QUANTITATIVE COMPARISON WITH CURRENT SOTA MODELS ON SIX WIDELY-USED DATASETS IN TERMS OF S-MEASURE (SM), F-MEASURE (FM), E-MEASURE (EM), AND MEAN ABSOLUTE ERROR (MAE). ↑ MEANS THAT THE LARGER THE NUMERICAL VALUE, THE BETTER THE MODEL, WHILE ↓ MEANS THE OPPOSITE. THE TOP-3 RESULTS ARE RESPECTIVELY MARKED IN RED, GREEN AND BLUE

Set	Metric	PCA 2018	PDNet 2018	CPFP 2019	DMRA 2019	S2MA 2020	A2dele 2020	D3Net 2020	DANet 2020	ICNet 2020	CoNet 2020	UCNet 2020	cmMS 2020	CMW 2020	ATSA 2020	BBSNet 2020	BTSNet 2021	DCF 2021	SPNet 2021	ASIF 2021	**Ours** 2022
NJUD	Sm↑	.877	.883	.878	.886	.894	.871	.892	.899	.894	.895	.897	.900	.903	.901	.919	.921	.919	.920	.889	.921
	Fm↑	.844	.832	.877	.872	.889	.874	.863	.871	.843	.872	.895	.897	.880	.893	.899	.901	.901	.899	.888	.903
	Em↑	.896	.890	.906	.908	.930	.897	.913	.926	.913	.927	.915	.922	.923	.921	.919	.926	.920	.920	.925	.930
	MAE↓	.059	.062	.053	.051	.053	.051	.047	.045	.052	.046	.043	.044	.046	.040	.037	.038	.039	.036	.047	.037
NLPR	Sm↑	.873	.835	.888	.899	.915	.898	.902	.920	.923	.908	.920	.915	.917	.907	.926	.930	.925	.926	.906	.933
	Fm↑	.794	.740	.822	.855	.902	.878	.857	.875	.908	.846	.901	.896	.872	.876	.878	.898	.899	.898	.888	.901
	Em↑	.916	.876	.924	.942	.953	.945	.943	.945	.952	.945	.953	.949	.941	.945	.949	.955	.958	.957	.944	.955
	MAE↓	.044	.064	.036	.031	.030	.028	.030	.027	.028	.031	.025	.027	.029	.028	.028	.021	.023	.024	.030	.022
SIP	Sm↑	.842	.783	.850	.806	.872	.829	.860	.875	.854	.858	.875	.872	.868	.864	.874	.882	.883	.880	.857	.884
	Fm↑	.824	.620	.818	.819	.849	.825	.835	.855	.791	.842	.876	.877	.851	.873	.874	.876	.872	.875	.859	.877
	Em↑	.898	.802	.899	.863	.911	.892	.902	.917	.900	.913	.913	.907	.909	.912	.915	.918	.912	.908	.896	.920
	MAE↓	.071	.166	.064	.085	.058	.070	.063	.054	.070	.063	.051	.058	.062	.058	.056	.052	.055	.055	.061	.051
STEREO	Sm↑	.880	.874	.871	886	.890	.879	.885	.892	.891	.908	.903	.895	.902	.897	.909	.905	.904	.905	.868	.910
	Fm↑	.845	.833	.827	.868	.882	.874	.855	.881	.847	.904	.899	.879	.867	.884	.886	.891	.895	.893	.893	.892
	Em↑	.905	.903	.897	.920	.932	.915	.920	930	.925	.937	.922	.927	.917	.919	.927	.935	.926	.924	.918	.939
	MAE↓	.061	.064	.054	.047	.051	.044	.046	.048	.046	.040	.039	.043	.044	.039	.041	.039	.041	.040	.049	.037
LFSD	Sm↑	.800	.845	.828	.847	.837	.834	.825	.845	.848	.862	.864	-	.866	.865	.856	.867	.862	.867	.814	.867
	Fm↑	.794	.824	.813	.849	.835	.832	.810	.846	.861	.859	.864	-	.871	.862	.850	.860	.860	.858	.858	.862
	Em↑	.846	.872	.867	.899	.873	-	.862	-	.887	-	.891	.896	-	-	.889	.896	.886	.890	-	.901
	MAE↓	.112	.109	.088	.075	.094	.077	.095	.083	.075	.071	.066	-	.067	.064	.074	.070	.060	.062	.089	.059
SSD	Sm↑	.843	.802	.807	.857	.868	-	.847	-	-	-	-	-	-	-	.870	-	.855	..852	.857	.872
	Fm↑	.786	.716	.725	.821	.848	-	.815	-	-	-	-	-	-	-	.832	-	.826	.814	.834	.850
	Em↑	.883	.813	.832	.892	.909	-	.888	-	-	-	-	-	-	-	.904	-	.898	.890	.884	.907
	MAE↓	.064	.115	.082	.058	.052	-	.058	-	-	-	-	-	-	-	.049	-	.053	.053	.056	.045

is trained on a composite training set, including 1,400 samples from NJUD [95] dataset and 650 samples from NLPR [96] dataset, and the input is resized to 352 × 352 resolution. The initial parameters of the feature encoding network are adopted from the pre-trained ResNet50 model [51]. The learning rate is set to 1e-4 for the Adam optimizer [103] and is later decayed by 10 at 60 epochs. We adopt cross-entropy loss for supervision.

D. Comparison With State-of-the-Arts

To demonstrate the effectiveness of the proposed method, we compare it with 19 **s**tate-**o**f-**t**he-**a**rt (SOTA) methods, *i.e.*, PCA18 [104], PDNet19 [105], CPFP19 [106], DMRA19 [40], S2MA20 [80], A2dele20 [107], D3Net21 [53], DANet20 [108], ICNet20 [72], cmMS20 [109], CMWNet20 [110], CoNet20 [111], UCNet20 [52], ATSA20 [112], BBSNet21 [70], BTSNet21 [64], DCF21 [113], SPNet21 [74], and ASIF21 [68].

The compared results are either reproduced by the released codes or saliency maps provided by authors, and the quantitative comparison results are shown in Table. II. Our method (see in Sec. III-C) performs the best on LFSD and NLPR datasets and shows competitive performance on STEREO, NJUD, and SIP datasets. More clearly, our method consistently outperforms all other compared SOTA methods in terms of the Fm metric. In the STEREO set, our method outperforms other competitors in terms of the Sm metric, *e.g.*, 0.910 (ours) v.s. 0.909 (BBSNet). In the LFSD set, our method improves 3.3% in terms of the Fm metric.

Fig. 8 further shows several visual comparisons of MaD with the latest representative models. From top to bottom, in rows #1, #2, and #5, we show three examples when image scenes with poor depths. Our method produces more reliable

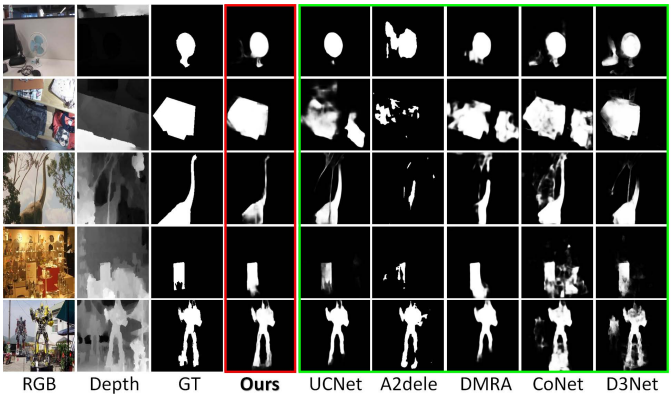


Fig. 8. Visual comparison between our method and several most representative SOTA models.

results, while other RGB-D saliency detection models fail to locate salient objects in images with low-quality depth. In row #7, we show an example with low contrast RGB, where it is challenging to locate all salient objects accurately. Our method locates all salient objects and segments them more accurately, generating sharper edges than other approaches. Moreover, in row #4 and row #8, both RGB and depth are of high quality, and our method generates the best result than any other SOTAs. We also show an example under complex conditions with fine-grained details in row #3 and row #6. Some approaches fail to complete detection, but our method can still perform well.

E. Component Evaluation

To validate the effectiveness of our method, we have conducted an extensive component evaluation, and the results have been shown in Table I. To enable a successful code running,

TABLE III
ABLATION STUDIES TOWARDS THE EFFECTIVENESS OF DTM AND MRM. THE BEST RESULTS ARE MARKED BY **BOLD** FONT

Datasets	NJUD				NLPR				SIP				STEREO				LFSD				SSD			
Metrics	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓
w/o DTM	.900	.882	.923	.053	.920	.891	.949	.027	.850	.847	.913	.079	.889	.856	.929	.024	.852	.850	.854	.080	.865	.840	.897	.050
DTM†	.912	.911	.927	.042	.927	.895	.953	.025	.865	.866	.918	.066	.900	.935	.875	.036	.863	.858	.859	.075	.869	.844	.903	.047
DTM††	.910	.899	.926	.047	.923	.894	.952	.024	.859	.858	.916	.071	.895	.868	.933	.031	.861	.855	.857	.077	.868	.842	.901	.048
w DTM	**.921**	**.903**	**.930**	**.037**	**.933**	**.901**	**.955**	**.022**	**.884**	**.877**	**.920**	**.051**	**.910**	**.892**	**.939**	**.037**	**.867**	**.862**	**.901**	**.059**	**.872**	**.850**	**.907**	**.045**
w/o MRM	.908	.875	.909	.047	.923	.884	.925	.029	.876	.860	.908	.057	.902	.881	.930	.042	.860	.843	.889	.068	.862	.841	.900	.052
MUL	.914	.893	.924	.040	.930	.897	.951	.023	.882	.871	.917	.054	.908	.888	.936	.038	.864	.858	.900	.060	.858	.847	.905	.048
CAT	.912	.890	.921	.044	.927	.891	.947	.025	.879	.863	.915	.055	.905	.883	.932	.041	.862	.850	.889	.064	.869	.842	.902	.050
w MRM	**.921**	**.903**	**.930**	**.037**	**.933**	**.901**	**.955**	**.022**	**.884**	**.877**	**.920**	**.051**	**.910**	**.892**	**.939**	**.037**	**.867**	**.862**	**.901**	**.059**	**.872**	**.850**	**.907**	**.045**

we have replaced those key components which need to be verified by simple operations, *e.g.*, the proposed MRM has been replaced by simple feature concatenation and convolution. We treat this replaced model as a baseline, and the qualitative result has been shown in the 1st column denoted by '**B**'.

Denoted by mark ①, the effectiveness of the proposed DTM (Sec. III-B) can be observed by comparing line 1 and line 2, where the baseline model equipped with DTM can achieve persistent performance gain, *e.g.*, 0.820 *vs.* 0.851 in terms of Fm in NJUD set.

Marked by ②, *i.e.*, lines 3-5 and 11, we can easily verify the effectiveness of the proposed MRM (Sec. III-C.1). In these cases, we have removed MRM from either modality-wise reasoning (denoted by MR), level-wise reasoning (characterized by LR), or both. Comparing models partially using (lines 3-4), the model using MRM (line 11) can steadily improve overall performance. Furthermore, the model that removed MRM (line 5) performs the worst. Notice that the model using MRM in Modality-wise Reasoning is slightly better than the model using MRM in Level-wise Reasoning, *e.g.*, the Fm metric has been improved from 0.899 to 0.901 in the NJUD set, showing the importance of the proposed Modality-wise Reasoning.

As is indicated by mark ③, lines 6-7 can well reflect the effectiveness of FE (Sec. III-C.1) and SF/DF (Sec. III-C.1) and the necessity of using Level-wise Reasoning can be confirmed by line 8. Compared with our complete model in line 11, both models either without using FE or SF/DF perform worse, where the Sm metric in the NLPR set has decreased respectively from 0.901→0.888 and 0.901→0.891. The reason is also quite apparent, *i.e.*, FE can integrate different modality features and obtain more delicate regions of interest, and SF/DF can achieve modality-aware and multi-level feature collection. Also, compared with line 11, line 8 illustrates the performance of a model which removes the Level-wise Reasoning, and, as expected, the performance drops significantly.

Highlighted by mark ④, we have verified the necessity of the proposed N-step MaF reasoning (line 9) and the effectiveness of using the last layer of the encoder as MaD's input (line 10). As shown in line 9 (we set the default iteration time as 3, and the corresponding ablation study will be conducted in Sec. IV-F.4), repeating MaF multiple times could improve performance as expected. Also, as denoted by line 10, the advantage of using the 5th level features against the 3rd level features as the MaD's input can be easily observed, e.g., 0.928 *vs.* 0.933 in terms of Sm in the NLPR set. The reason is that the 5th level features contain more semantic information, which can be helpful to precisely locate salient objects. In contrast, the 3rd level features have redundant information, hindering modality-aware fusion.

F. Ablation Studies

We have provided comprehensive ablation studies to further evaluate the contribution of each key component in our method. Specifically, we investigate 1) the importance of DTM, 2) the effectiveness of MRM, 3) the influence of iterations of MaD, and 4) the number of sequential 1D convolutions. We change one component each time and retrain variants with the same hyperparameters and training settings.

1) Importance of DTM: To validate the effectiveness of the proposed DTM (Sec. III-B), we set up three experiments with the same parameter settings. Specifically, 'DTM†' means removing all data flows between sequential DTMs in horizontal direction (*i.e.*, f_{dr}^{i} in Fig. 4), while 'DTM††' denotes removing all vertical data flows which output depth information to RGB stream. We use 'w/o DTM' to represent without using DTM. As shown in Table III, when no interaction exists between RGB and D branches, the model performs the worst. Also, though using only partial interactions between DTM units can achieve some performance gain (about 2.6% improvement in terms of Sm metric), the complete interaction version has exhibited an undeniable advantage, *e.g.*, the Fm metric can be improved from 0.875 to 0.892 and 0.868 to 0.892 respectively in the STEREO set. Such comparative experiments show that the DTM is crucial for downstream Modality-aware Reasoning. Our proposed DTM can mine high-quality and practical D and use them to promote their RGB counterparts.

2) Effectiveness of MRM: In the proposed framework, the MRM (Sec. III-C.1) is adopted to learn inter-modality relationships, where the learned relationship is used to guide the fusion process between RGB and D. To validate its effectiveness, we have tried to delete this MRM (this model has also been reported in line 5 of Table I), denoted as 'w/o MRM'.

Besides, we have also compared two other plain feature fusion strategies with our MRM, *i.e.*, performing inter modality fusion via either element-wise multiplication (denoted by 'MUL') or simple concatenation with convolution (denoted by 'CAT'). As shown in Table III, comparing 'w/o MRM' with

TABLE IV

QUANTITATIVE RESULTS OF THE NUMBER OF MAD ITERATIONS (N). THE BEST RESULT IS HIGHLIGHTED BY **BOLD** FONT

Set	NJUD				NLPR				SIP				STEREO				LFSD				SSD			
Metric	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓
N=1	.918	.899	.924	.038	.932	.898	.948	.024	.882	.873	.909	.053	.907	.890	.928	.039	.863	.859	.893	.070	.864	.838	.901	.051
N=2	.919	.901	.926	.036	.932	.899	.949	.024	.882	.875	.912	.052	.908	**.892**	.931	.038	.866	.860	.896	.070	.867	.841	.902	.048
N=3	**.921**	**.903**	**.930**	**.037**	**.933**	**.901**	**.955**	**.022**	**.884**	**.877**	**.920**	**.051**	**.910**	**.892**	**.939**	**.037**	**.867**	**.862**	**.901**	**.059**	**.872**	**.850**	**.907**	**.045**
N=4	.919	.902	.927	**.037**	.931	.897	.952	.023	.883	.876	.915	.052	.907	.891	.936	.038	.866	.861	.899	.063	.870	.845	.905	.047

TABLE V

ABLATION STUDY REGARDING THE NUMBER (i) OF SEQUENTIAL 1D CONVOLUTIONS (SEC. III-C.1)

Set	NJUD				NLPR				SIP				STEREO				LFSD				SSD			
Metric	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓
i=1	**.921**	**.903**	**.930**	**.037**	**.933**	.901	**.955**	**.022**	**.884**	**.877**	**.920**	**.051**	**.910**	**.892**	**.939**	**.037**	**.867**	**.862**	**.901**	**.059**	**.872**	**.850**	**.907**	**.045**
i=2	.919	.902	.928	.036	.930	**.903**	.952	.023	.882	.876	.927	.052	.909	**.892**	.936	.036	.866	.860	.898	.060	.868	.849	.903	.046
i=3	.918	.898	.925	.037	.928	.898	.949	.025	.881	.875	.925	.053	.907	.890	.934	.039	.863	.857	.895	.061	.867	.848	.905	.048
i=4	.916	.900	.921	.039	.927	.897	.946	.025	.878	.872	.921	.055	.903	.876	.933	.040	.860	.855	.893	.063	.864	.845	.901	.051

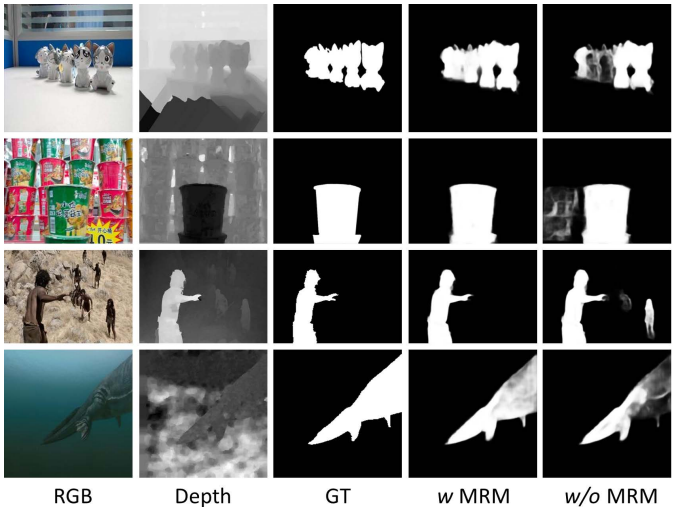


Fig. 9. Qualitative demonstration of **m**odality **r**elationship **m**odule (MRM).

our full model ('w MRM'), there is an average 2% margin in terms of the Fm metric. Comparing 'MUL' and 'CAT' with our full model, we can see that our MRM outperforms them clearly, *e.g.*, 0.914 *vs.* 0.921 and 0.912 *vs.* 0.921 in terms of Sm over NLPR set. The results demonstrate the effectiveness of extracting high-level feature representations using graph relation among different modalities. The reason is that graph structure can obtain a harmonious degree between different modalities to learn the complementary relationships. For a better reading, we have also provided some qualitative demonstrations in Fig. 9.

3) Validity of Modality-Aware Fusion (MaF): To further verify the advantages of the proposed "modality-aware fusion", we have conducted a quantitative test to see if the claimed new fusion brings our performance gain. In the experiment, we propose to investigate the absolute performance gain obtained by the fusion module. Our rationale is that a more powerful fusion shall gain more performance against low-level saliency. Here we have selected 3 most representative SOTA models (*e.g.*, **SPNet21** [74], **ATSA20** [112], and **CMW20** [110]), where their RGB saliency and D saliency are respectively obtained by averaging each sub-stream's side-output quantitative scores. The results can be seen in Table VI.

As shown in the table, both low-level saliency (RGB saliency and D saliency) cues in our model have no clear advantages over other models. However, as suggested by the "Numeric Gain", our fusion process has performed the best. Notice that, since our fusion has solely focused on the learning modality relationships with a much simpler implementation, it is reasonable to infer that the proposed modality-aware fusion is very effective. Besides, the Numeric Gains achieved by the other three compared models are pretty limited because these methods have failed to be completely modality-aware.

4) Influence of MaD Iterations: We have conducted an ablation study regarding the iterations N (Sec. III-C), and the detailed results can be found in Table IV. Limited by GPU storage, we only chose $N = 1, 2, 3, 4$, in which $N = \mathrm{i}$ means to iterate the whole MaD for i times. As shown in Table IV, the overall performance of our method is moderately sensitive to the choice of N, in which the overall performance via $N = 3$ exhibits the best performance in general, and explicit performance degradation can be found when we assign $N = 1$. Meanwhile, when assigning N to 4, the performance has also decreased, *e.g.*, 0.933 *vs.* 0.931 in terms of the Sm metric on the NLPR set, and thus we set $N = 3$ as the optical choice to strike the trade-off between accuracy and efficiency.

Further, we have included qualitative illustrations of the modality-aware (ours) and modality-unware (S2MA [80]) method in the RGB-D fusion process, which can be seen in Fig. 10. Expressly, we have provided the last three MaF output results (because we adopt three MaFs in our network to balance the efficiency and over-smoothing problem), *e.g.*, MaF-1, MaF-2, and MaF-3 (the MaF-3 is utilized to produce final saliency maps). And also, we have illustrated the last three fusion processing results of S2MA, denoted by D1, D2, and D3 (the D3 is utilized to produce final saliency maps). As we can see, in almost all scenes of low contrast (row 1), complex background (row 2), and simple objects (row 3), our MaF-3 can obtain superior visual results. The main reason is that the modality-unaware method has "equally" considered depth and RGB for all image scenes. As a result, when one

TABLE VI

QUANTITATIVE VERIFICATIONS TOWARDS THE ADVANTAGES OF BEING COMPLETE MODALITY-AWARE. HERE WE COMPARE OUR METHOD WITH THE OTHER 3 REPRESENTATIVE SOTA MODELS. 'NUMBER GAIN' CAN BE COMPUTED BY: {FUSED_SAL−MAX(RGB_SAL, DEPTH_SAL)}, WHERE RGB_SAL AND DEPTH_SAL ARE THE LOW-LEVEL SALIENCY CUES, WHICH CAN BE COMPUTED BY AVERAGING THE NUMERIC SCORES OF SIDE-OUTPUTS RESPECTIVELY FROM RGB BRANCH AND DEPTH BRANCH OF THE CONSIDERED SOTA MODEL

SPNet21 [74]	NJUD [95]			STEREO [97]			SIP [98]		
Metrics	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓
RGB_sal	0.901	0.885	0.042	0.908	0.890	0.041	0.835	0.820	0.093
Depth_sal	0.858	0.834	0.064	0.735	0.695	0.120	0.721	0.712	0.145
Fused_sal	0.920	0.899	0.036	0.905	0.893	0.040	0.867	0.858	0.062
Numeric Gain	0.019	0.014	-0.006	-0.003	0.003	-0.001	0.032	0.038	-0.031
ATSA20 [112]	NJUD [95]			STEREO [97]			SIP [98]		
Metrics	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓
RGB_sal	0.886	0.895	0.043	0.892	0.875	0.042	0.834	0.839	0.069
Depth_sal	0.855	0.867	0.051	0.873	0.760	0.081	0.821	0.817	0.085
Fused_sal	0.901	0.893	0.040	0.897	0.884	0.039	0.864	0.873	0.058
Numeric Gain	0.015	-0.002	-0.003	0.005	0.009	-0.003	0.030	0.034	-0.011
CMW20 [110]	NJUD [95]			STEREO [97]			SIP [98]		
Metrics	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓
RGB_sal	0.907	0.872	0.053	0.904	0.862	0.049	0.843	0.823	0.071
Depth_sal	0.873	0.855	0.057	0.863	0.820	0.058	0.822	0.815	0.083
Fused_sal	0.903	0.880	0.046	0.902	0.867	0.044	0.868	0.851	0.062
Numeric Gain	-0.004	0.008	-0.007	-0.002	0.005	-0.005	0.025	0.028	-0.009
Ours	NJUD [95]			STEREO [97]			SIP [98]		
Metrics	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓
RGB_sal	0.896	0.884	0.046	0.903	0.880	0.046	0.846	0.833	0.090
Depth_sal	0.842	0.851	0.069	0.749	0.703	0.099	0.823	0.775	0.116
Fused_sal	0.921	0.903	0.037	0.910	0.892	0.037	0.884	0.877	0.051
Numeric Gain	**0.025**	**0.019**	**-0.009**	**0.007**	**0.012**	**-0.009**	**0.038**	**0.044**	**-0.039**

TABLE VII

DETAILED AVERAGED TIME COST FOR A SINGLE IMAGE. THIS RESULT WAS OBTAINED ON A PC WITH AN INTEL(R) XEON(R) CPU, NVIDIA GTX2080 GPU (WITH 8G RAM) AND 32G RAM. THIS EXPERIMENT WAS CARRIED OUT ON THE SSD SET

Main Steps	Seconds
Key Comp. 1: Depth Transfer Module (Sec. III-B)	**0.0003**s
Key Comp. 2: Modality-aware Fusion × 3 (Sec. III-C)	**0.0172**s
(1) *Modality-wise Reasoning*	0.0098s
MRM (Sec. III-C1)	0.0065s
FE (Sec. III-C2)	0.0031s
SF (Sec. III-C3)	0.0002s
(2) *Level-wise Reasoning*	0.0074s
MRM (Sec. III-C1)	0.0041s
FE (Sec. III-C2)	0.0031s
SF (Sec. III-C3)	0.0002s
Key Comp. 3: Saliency Decoder (Sec. III-D)	**0.0008**s
Total	**0.0183**s

or both of the two modalities are of poor quality, the one with more inferior quality will bring a negative impact.

We have also included the computational cost of each component in Table VII. We can see that the primary time computation of the network lies in the Key Component 2 — Modality-aware Fusion, which takes almost 95% of the total time. Notice that each sub-part in Key Component 2 has a reasonable computation cost.

5) Numbers of Sequential 1D Convolutions: To verify the effectiveness of the adopted 1D Convolutions in the learning modality relationship (Sec. III-C.1), we have conducted an extensive ablation study regarding the number of 1D Convolutions from 1 group to 4 groups (two sequential 1D Convolutions are regarded as a group, indicated by i). From detailed results in Table. V we can see, with the increased groups of 1D convolutions, the performance has declined, *e.g.*, when increasing i from 1 to 4, the Sm metric in the NJUD set has decreased from 0.921→0.916. The reason is quite similar to the phenomenon in fully convolution layers — too many sequential non-linear mappings could lead to the learned model overfitting. Thus, we have chosen $i = 1$ as our default setting.

6) Applications of MaD to Other RGB-D SOTA Models: We have also tried to apply our method to other RGB-D SOTA models, as shown in Table VIII. In this experiment, we have newly replaced the decoding fusion strategies of several other RGB-D methods (BBSNet [70], SPNet [74], and DCF [113]) with our modality-aware decoder. We find that the three SOTA methods equipped with our modality-aware decoder can achieve better results, suggesting the relatively generic nature of our proposed method.

TABLE VIII

ABLATION STUDY ON OTHER RGB-D METHODS BY REPLACING THE DECODING FUSION STRATEGIES WITH OUR **M**ODALITY-**A**WARE **D**ECODER (MAD)

Dataset	NJUD [95]			STEREO [97]			LFSD [98]		
Metrics	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓	Sm↑	Fm↑	MAE↓
BBSNet [70]	0.919	0.899	0.037	0.909	0.886	0.041	0.856	0.850	0.074
BBSNet+MaD	0.925	0.910	0.035	0.915	0.892	0.038	0.858	0.855	0.072
SPNet [74]	0.920	0.899	0.036	0.905	0.893	0.040	0.867	0.858	0.062
SPNet+MaD	0.927	0.912	0.034	0.917	0.906	0.036	0.872	0.867	0.059
DCF [113]	0.919	0.901	0.039	0.904	0.895	0.041	0.862	0.860	0.060
DCF+MaD	0.922	0.908	0.037	0.908	0.900	0.037	0.878	0.867	0.056

G. Limitations

We demonstrate some failure cases in Fig. 11. Usually, our method still faces two challenges: 1) salient objects with varying subparts, *e.g.*, the 1st row of Fig. 11, and 2) image scenes with multiple salient objects, *e.g.*, the 2nd row of Fig. 11. In cases with high-quality depth yet salient objects with significant color differences between their subparts, our method fails to detect them accurately and mistakenly highlights some non-salient details. In the bottom row, 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 are usually localized in different depth layers, which mistakenly causes our method to treat salient objects far from the camera as distractions.

H. Visualized Comparison Between Modality-Aware and Modality-Unaware

In Fig 12, we have compared our method with some modality-unaware methods (UCNet20 [52], D3Net21 [53], ICNet20 [72], CoNet20 [111], BTSNet21 [64], and S2MA20 [80]) regarding four types of RGB and Depth combinations (see subfig-A to subfig-B). As shown in subfig-A of Fig 12, when both RGB and depth are of high

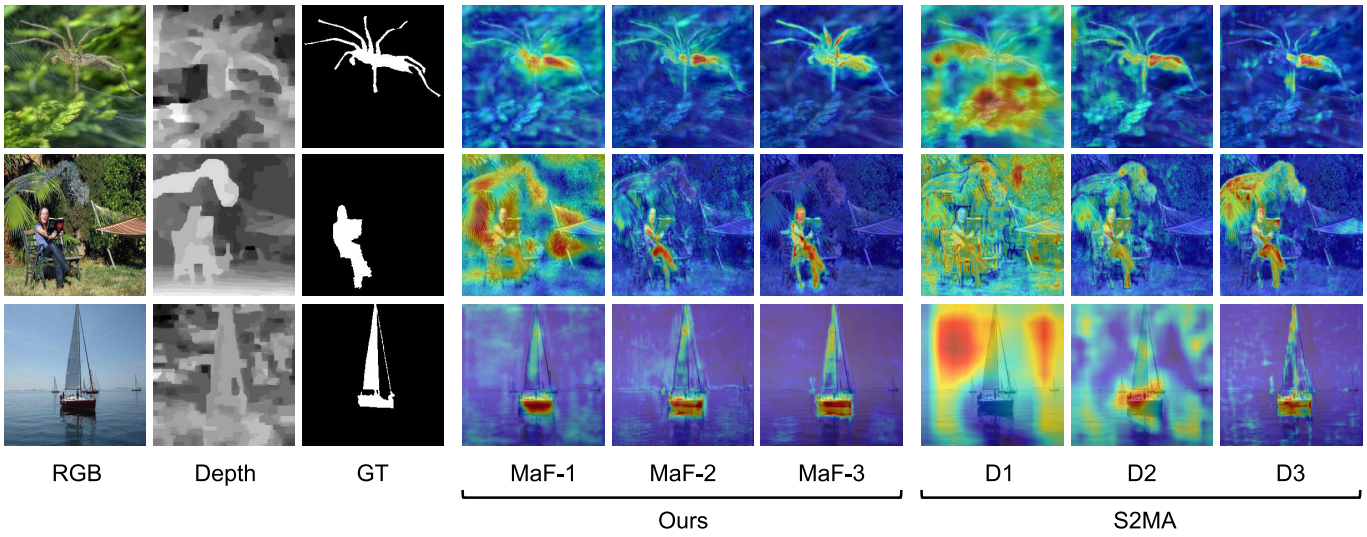


Fig. 10. Visualizations of modality-aware (ours) and modality-unware (S2MA [80]) methods in terms of the last three fusion processing results in decoding phase. MaF denotes Modality-aware Fusion.

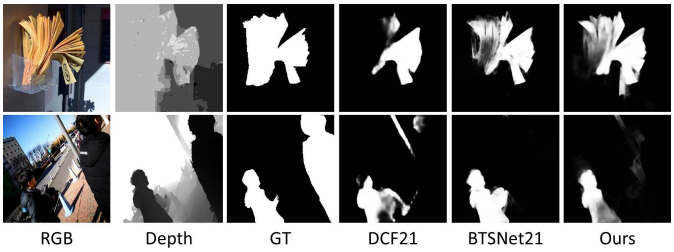


Fig. 11. Demonstration of some representative failure cases.

TABLE IX
MODEL SIZE AND RUNNING TIME COMPARISONS

Models	Sizes↓	FPS↑	NJUD				NLPR			
			Sm↑	Fm↑	Em↑	MAE↓	Sm↑	Fm↑	Em↑	MAE↓
D3Net [53]	530MB	32	.892	.863	.913	.047	.902	.857	.943	.030
UCNet [52]	119MB	42	.871	.874	.897	.051	.898	.878	.945	.028
CCAF [90]	547MB	36.4	.910	.898	.920	.037	.922	.882	.952	.026
S2MA [80]	352MB	26	.894	.889	**.930**	.053	.915	.902	.953	.030
CPFP [106]	375MB	7	.878	.844	.906	.059	.835	.740	.924	.064
ATSA [112]	123MB	29	.901	.893	.921	.040	.907	.876	.945	.028
A2dele [107]	**57MB**	**120**	.919	.899	.919	.037	.926	.878	.949	.028
cmMS [109]	430MB	15	.900	.897	.922	.044	.915	.896	.949	.027
Ours (B+DE)	306MB	46	.891	.825	.881	.058	.902	.890	.885	.031
Ours	310MB	52	**.921**	**.903**	**.930**	**.037**	**.933**	**.901**	**.955**	**.022**

quality (denoted by '+'), modality-aware and modality-unaware fusion-based methods can simultaneously obtain satisfying results in terms of scenes of clear boundary and multiple objects (though our modality-aware fusion slightly outperforms modality-unaware fusion), since both RGB and depth can positively contribute to the results. However, when one of the RGB and depth is low-quality (see subfig-B and subfig-C), the results of modality-unaware fusion-based methods degrade. The reason is that modality-unaware fusion-based methods do not appropriately learn the complemental relationship between RGB and depth, and they merely integrate the two-modality feature slices in a local manner. In practice, when one modality dominates (the higher-quality one) the fusion, the other (the lower-quality one) may hinder the fusion and bring about negative effects. Nevertheless, these methods treat the two modalities equally, ignoring the harmony degree (what we call **'modality-aware'**) between RGB and depth, which determines the degree of complementarity. Conversely, our modality-aware method can adaptively bias to the appropriate modality guided by the learned relationships between RGB and depth. Therefore, in cases of one modality dominating, our method is optimal and has a higher performance ceiling. In subfig-D, in more challenging cases when objects are occluded and small, and depth maps are fuzzy, our modality-aware fusion can still outperform those modality-unaware fusions since though both two RGB and depth are low-quality, our method can still appropriately learn the complemental relationship between RGB and depth and bias to the appropriate modality, which demonstrates the robustness of modality-aware fusion strategy.

In summary, our modality-aware fusion strategy can surpass most of the existing modality-unaware fusion and will enlighten future work regarding how to fully and appropriately learn the relationships during cross-modality fusion.

I. Discussion of Our Performance Gain

Because we have adopted multiple sequential MaF modules to learn inter-modality relationships, our model size is relatively large. In Table IX, we have conducted a model size comparison, where our model lies in the middle level among all compared methods. To verify whether our performance gain is brought by additional model size, we have increased the baseline model size by using additional decoder layers, ensuring a fair comparison in model size. This modified baseline model has been denoted by 'B+DE'. The results have shown that increasing the model's learning capacity cannot ensure a corresponding performance gain, even if the number of parameters is competitively large. By comparing

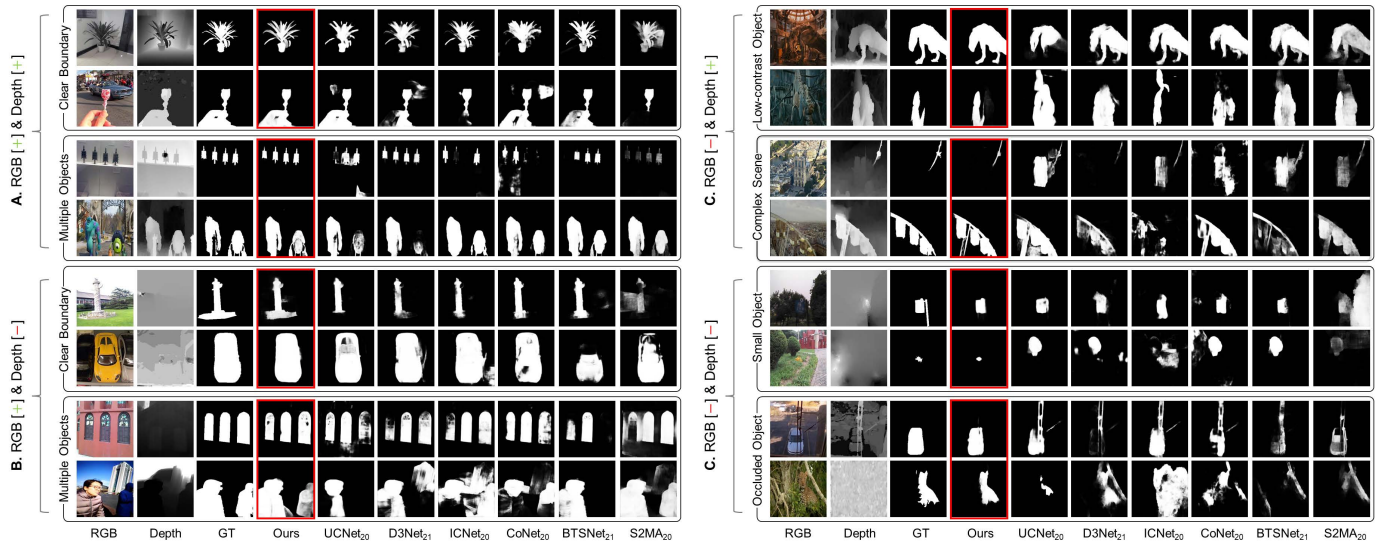


Fig. 12. Visualized comparison among our method and some modality-unaware methods (UCNet20 [52], D3Net21 [53], ICNet20 [72], CoNet20 [111], BTSNet21 [64], and S2MA20 [80]). '+' denotes high quality, and '-' denotes low quality.

the modified baseline with our model, we can easily notice a significant performance margin, *e.g.*, the Sm metric of NLPR has been increased from 0.902 to 0.933, which is a shred of solid evidence to show that the additional model size does not simply bring our performance gain.

V. Conclusion and Future Works

In this paper, we have proposed a novel modality-aware decoder to learn the relationship between different modalities. The learned inter-modality relationship is used to guide RGB-D saliency fusion. The essential technical contribution is a novel idea to enable the RGB-D fusion process to be modality-aware. Thus our fusion enables a significant performance improvement without fancy network design. Our key idea can also inspire other multi-modality-related fusion works, where the usage of intermodality relationships is beneficial in achieving better complementary status between different modalities. We have also conducted an extensive comparison and component evaluation, where the quantitative comparison has confirmed our performance gain, and the quantitative component evaluation has verified the effectiveness of each significant component adopted in our approach. We have also released our codes and results, which can potentially benefit our research community in the future. In the near future, we are particularly interested in reducing the model size without degrading model performance. We plan to devise more efficient graph-based operations to substitute current plain 1D convolutions, *i.e.*, adopting a more appropriate dynamic adjacent matrix to fit our RGB-D ISOD task.

References

[1] H. Zhai, S. Lai, H. Jin, X. Qian, and T. Mei, "Deep transfer hashing for image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 742–753, Feb. 2020.

[2] S. Kan, Y. Cen, Y. Cen, M. Vladimir, Y. Li, and Z. He, "Zero-shot learning to index on semantic trees for scalable image retrieval," *IEEE Trans. Image Process.*, vol. 30, pp. 501–516, 2021.

[3] S. Sharma, V. Gupta, and M. Juneja, "A novel unsupervised multiple feature hashing for image retrieval and indexing (MFHIRI)," *J. Vis. Commun. Image Represent.*, vol. 84, Apr. 2022, Art. no. 103467.

[4] Y. Wang, R. Zhao, L. Liang, X. Zheng, Y. Cen, and S. Kan, "Block-based image matching for image retrieval," *J. Vis. Commun. Image Represent.*, vol. 74, Jan. 2021, Art. no. 102998.

[5] L. Jiang, M. Xu, X. Wang, and L. Sigal, "Saliency-guided image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 16504–16513.

[6] K.-K. Babua and S.-R. Dubeyb, "CDGAN: Cyclic discriminative generative adversarial networks for image-to-image transformation," *J. Vis. Commun. Image Represent.*, vol. 82, Jan. 2022, Art. no. 103382.

[7] A. Raghunandan, P. Raghav, and H. V. R. Aradhya, "Object detection algorithms for video surveillance applications," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, 2018, pp. 563–568.

[8] Z.-M. Chen, X. Jin, B.-R. Zhao, X. Zhang, and Y. Guo, "HCE: Hierarchical context embedding for region-based object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 6917–6929, 2021.

[9] Y. Liu, Q. Minglang, M. Xu, B. Li, W. Hu, and A. Borji, "Learning to predict salient faces: A novel visual-audio saliency model," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12365, 2020, pp. 413–429.

[10] M. Xu, Y. Ren, Z. Wang, J. Liu, and X. Tao, "Saliency detection in face videos: A data-driven approach," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1335–1349, Jun. 2018.

[11] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for HEVC-MSP," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 155–170, Jan. 2018.

[12] K. B. Girum, G. Crhange, and A. Lalande, "Learning with context feedback loop for robust medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 6, pp. 1542–1554, Jun. 2021.

[13] J. Zhang, Y. Xie, Y. Wang, and Y. Xia, "Inter-slice context residual learning for 3D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 661–672, Feb. 2021.

[14] G. Lu, X. Zhang, W. Ouyang, D. Xu, L. Chen, and Z. Gao, "Deep non-local Kalman network for video compression artifact reduction," *IEEE Trans. Image Process.*, vol. 29, pp. 1725–1737, 2020.

[15] R. Yang, M. Xu, Z. Wang, Y. Duan, and X. Tao, "Saliency-guided complexity control for HEVC decoding," *IEEE Trans. Broadcast.*, vol. 64, no. 4, pp. 865–882, Dec. 2018.

[16] Q. Guo, W. Feng, R. Gao, Y. Liu, and S. Wang, "Exploring the effects of blur and deblurring to visual object tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 1812–1824, 2021.

[17] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1571–1580.

[18] J. Li, Z. Pan, Q. Liu, Y. Cui, and Y. Sun, "Complementarity-aware attention network for salient object detection," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 873–886, Feb. 2022.

[19] S. Chen, B. Wang, X. Tan, and X. Hu, "Embedding attention and residual network for accurate salient object detection," *IEEE Trans. Cybern.*, vol. 50, no. 5, pp. 2050–2062, May 2020.

[20] Y. Liu, M.-M. Cheng, X.-Y. Zhang, G.-Y. Nie, and M. Wang, "DNA: Deeply supervised nonlinear aggregation for salient object detection," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6131–6142, Jul. 2022.

[21] H. Li, G. Li, B. Yang, G. Chen, L. Lin, and Y. Yu, "Depthwise nonlocal module for fast salient object detection using a single thread," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 6188–6199, Dec. 2021.

[22] Y. Zhou, S. Huo, W. Xiang, C. Hou, and S. Kung, "Semi-supervised salient object detection using a linear feedback control system model," *IEEE Trans. Cybern.*, vol. 49, no. 4, pp. 1173–1185, Apr. 2018.

[23] J.-J. Liu, Q. Hou, and M.-M. Cheng, "Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton," *IEEE Trans. Image Process.*, vol. 29, pp. 8652–8667, 2020.

[24] S. Yang, W. Lin, G. Lin, Q. Jiang, and Z. Liu, "Progressive self-guided loss for salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 8426–8438, 2021.

[25] J. Li, J. Su, C. Xia, M. Ma, and Y. Tian, "Salient object detection with purificatory mechanism and structural similarity loss," *IEEE Trans. Image Process.*, vol. 30, pp. 6855–6868, 2021.

[26] Z. Wu, L. Su, and Q. Huang, "Decomposition and completion network for salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 6226–6239, 2021.

[27] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "SAM-Net: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3804–3814, 2021.

[28] M. Xu, L. Yang, X. Tao, Y. Duan, and Z. Wang, "Saliency prediction on omnidirectional image with generative adversarial imitation learning," *IEEE Trans. Image Process.*, vol. 30, pp. 2087–2102, 2021.

[29] M. Xu, L. Jiang, Z. Wang, and L. Sigal, "DeepVS2.0: A saliency-structured deep learning method for predicting dynamic visual attention," *Int. J. Comput. Vis.*, vol. 129, pp. 203–224, Jan. 2021.

[30] Y. Wu, Z. Liu, and X. Zhou, "Saliency detection using adversarial learning networks," *J. Vis. Commun. Image Represent.*, vol. 67, Feb. 2020, Art. no. 102761.

[31] X. Zhou *et al.*, "Edge-guided recurrent positioning network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, early access, Apr. 13, 2022, doi: 10.1109/TCYB.2022.3163152.

[32] Z. Zhou, Y. Guo, J. Huang, M. Dai, M. Deng, and Q. Yu, "Superpixel attention guided network for accurate and real-time salient object detection," *Multimedia Tools Appl.*, vol. 2022, pp. 1–24, Apr. 2022.

[33] V.-K. Singh and N. Kumar, "CoBRa: Convex hull based random walks for salient object detection," *Multimedia Tools Appl.*, vol. 81, pp. 30283–30303, Apr. 2022.

[34] 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, 2020.

[35] 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, 2020.

[36] 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.*, vol. 30, pp. 7012–7024, 2021.

[37] 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.

[38] Z. Liu, Y. Wang, Z. Zhang, and Y. Tan, "BGRDNet: RGB-D salient object detection with a bidirectional gated recurrent decoding network," *Multimedia Tools Appl.*, vol. 81, pp. 25519–25539, Mar. 2022.

[39] H. Xu, J. Xu, and W. Xu, "Survey of 3D modeling using depth cameras," *Virtual Reality Intell. Hardw.*, vol. 1, no. 5, pp. 483–499, 2019.

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

[41] 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.

[42] 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.

[43] J. Guo, T. Ren, and J. Bei, "Salient object detection for RGB-D image via saliency evolution," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2016, pp. 1–6.

[44] 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, Apr. 2016.

[45] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y.-C. Gu, and M.-M. Cheng, "MobileSal: Extremely efficient RGB-D salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 13, 2021, doi: 10.1109/TPAMI.2021.3134684.

[46] F. Wang, J. Pan, S. Xu, and J. Tang, "Learning discriminative cross-modality features for RGB-D saliency detection," *IEEE Trans. Image Process.*, vol. 31, pp. 1285–1297, 2022.

[47] 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.

[48] H. Wen *et al.*, "Dynamic selective network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 9179–9192, 2021.

[49] C. Zhang *et al.*, "Cross-modality discrepant interaction network for RGB-D salient object detection," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 2094–2102.

[50] X. Zhou *et al.*, "FANet: Feature aggregation network for RGBD saliency detection," *Signal Process., Image Commun.*, vol. 102, Mar. 2022, Art. no. 116591.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

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

[53] 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.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[54] C. Zhu and G. Li, "A multilayer backpropagation saliency detection algorithm and its applications," *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 25181–25197, 2018.

[55] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2017, pp. 1509–1515.

[56] L. Wu, Z. Liu, H. Song, and O. Le Meur, "RGBD co-saliency detection via multiple kernel boosting and fusion," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 21185–21199, Jan. 2018.

[57] H. Song, Z. Liu, Y. Xie, L. Wu, and M. Huang, "RGBD co-saliency detection via bagging-based clustering," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1722–1726, Dec. 2016.

[58] 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*, Jun. 2015, pp. 25–32.

[59] 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, 2021.

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

[61] J. Zhang *et al.*, "Uncertainty inspired RGB-D saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5761–5779, Sep. 2022.

[62] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1407–1417.

[63] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 3138–3147.

[64] W. Zhang, Y. Jiang, K. Fu, and Q. Zhao, "BTS-Net: Bi-directional transfer-and-selection network for RGB-D salient object detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2021, pp. 1–6.

[65] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3049–3059.

[66] X. Zhou *et al.*, "Dense attention-guided cascaded network for salient object detection of strip steel surface defects," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[67] X. Zhou, H. Fang, X. Fei, R. Shi, and J. Zhang, "Edge-aware multi-level interactive network for salient object detection of strip steel surface defects," *IEEE Access*, vol. 9, pp. 149465–149476, 2021.

[68] 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.

[69] 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.

[70] Y. Zhai *et al.*, "Bifurcated backbone strategy for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 8727–8742, 2021.

[71] X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, and H. Qin, "Data-level recombination and lightweight fusion scheme for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 458–471, 2021.

[72] 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, 2020.

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

[74] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving RGB-D saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 4681–4691.

[75] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 369–378.

[76] B. He, X. Yang, Z. Wu, H. Chen, S.-N. Lim, and A. Shrivastava, "GTA: Global temporal attention for video action understanding," 2020, *arXiv:2012.08510*.

[77] K. Choromanski *et al.*, "Rethinking attention with performers," 2020, *arXiv:2009.14794*.

[78] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel 'squeeze and excitation' blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Feb. 2019.

[79] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3668–3677.

[80] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13753–13762.

[81] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3528–3542, 2021.

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

[83] H. Xia and X. Gao, "Multi-scale mixed dense graph convolution network for skeleton-based action recognition," *IEEE Access*, vol. 9, pp. 36475–36484, 2021.

[84] M. Xu, P. Fu, B. Liu, and J. Li, "Multi-stream attention-aware graph convolution network for video salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 4183–4197, 2021.

[85] A. Luo, X. Li, F. Yang, Z. Jiao, and S. Lyu, "Cascade graph neural networks for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 346–364.

[86] B. Jiang, X. Jiang, J. Tang, B. Luo, and S. Huang, "Multiple graph convolutional networks for co-saliency detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 332–337.

[87] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "TriTransNet: RGB-D salient object detection with a triplet transformer embedding network," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4481–4490.

[88] X. Wang *et al.*, "Boosting RGB-D saliency detection by leveraging unlabeled RGB images," *IEEE Trans. Image Process.*, vol. 31, pp. 1107–1119, 2022.

[89] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "CDNet: Complementary depth network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3376–3390, 2021.

[90] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images," *IEEE Trans. Multimedia*, vol. 24, pp. 2192–2204, 2022.

[91] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 433–442.

[92] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1025–1035.

[93] J. Gao, J. Gao, X. Ying, M. Lu, and J. Wang, "Higher-order interaction Goes neural: A substructure assembling graph attention network for graph classification," *IEEE Trans. Knowl. Data Eng.*, early access, Aug. 18, 2021, doi: 10.1109/TKDE.2021.3105544.

[94] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3902–3911.

[95] 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.

[96] H. Peng, L. Bing, 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.

[97] 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.

[98] 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. 2016.

[99] G. Li and C. Zhu, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3008–3014.

[100] 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. 4558–4567.

[101] 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.

[102] 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. 1–66.

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

[104] 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.

[105] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2019, pp. 199–204.

[106] 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.

[107] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 9057–9066.

[108] 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. Eur. Conf. Comput. Vis.*, 2020, pp. 646–662.

[109] C. Li, R. Cong, Y. Piao, and Q. Xu, "RGB-D salient object detection with cross-modality modulation and selection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 225–241.

[110] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 665–681.

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

[112] M. Zhang, S. X. Fei, J. Liu, S. Xu, and H. Lu, "Asymmetric two-stream architecture for accurate RGB-D saliency detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 374–390.

[113] W. Ji *et al.*, "Calibrated RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9466–9476.