Data-Level Recombination and Lightweight Fusion Scheme for RGB-D Salient Object Detection

Xuehao Wang[1] Shuai Li[1] Chenglizhao Chen[1,2*] Yuming Fang[3] Aimin Hao[1] Hong Qin[4]
[1]Qingdao Research Institute & State Key Laboratory of VRTS, Beihang University
[2]Qingdao University [3]Jiangxi University of Finance and Economics [4]Stony Brook University

***Abstract*—Existing RGB-D salient object detection methods treat depth information as an independent component to complement RGB and widely follow the bistream parallel network architecture. To selectively fuse the CNN features extracted from both RGB and depth as a final result, the state-of-the-art (SOTA) bistream networks usually consist of two independent subbranches: one subbranch is used for RGB saliency, and the other aims for depth saliency. However, depth saliency is persistently inferior to the RGB saliency because the RGB component is intrinsically more informative than the depth component. The bistream architecture easily biases its subsequent fusion procedure to the RGB subbranch, leading to a performance bottleneck. In this paper, we propose a novel data-level recombination strategy to fuse RGB with D (depth) before deep feature extraction, where we cyclically convert the original 4-dimensional RGB-D into DGB, RDB and RGD. Then, a newly lightweight designed triple-stream network is applied over these novel formulated data to achieve an optimal channel-wise complementary fusion status between the RGB and D, achieving a new SOTA performance.**

***Index Terms*—RGB-D Saliency Detection, Data-level Fusion, Lightweight Designed Triple-stream Network.**

I. Introduction and Motivation

Salient object detection (SOD) aims to separate the most visually distinctive objects from nonsalient nearby surroundings [1]–[3]. As a widely used preprocessing tool, the SOD related down-stream applications include various computer vision tasks, such as object detection [4], [5], image expression and enhancement [6], [7], image retrieval [8], [9], image compression [10], [11], image retargeting [12], [13], visual tracking [14], [15], video saliency [16]–[20] and video segmentation [21]–[24].

Different from the conventional salient object detection methods [25]–[30] using solely RGB information, the RGB-D salient object detection methods [31]–[37] have achieved significant performance improvements due to the newly available depth information (see Fig. 1). In fact, we may easily obtain a high-quality saliency map over the depth channel if it satisfies the following aspects: 1) the depth information is correctly sensed by the depth-sensing equipment; 2) the salient object is located at a different depth layer with respect to its nonsalient nearby surroundings.

The conventional SOTA RGB-D salient object detection methods [38] have adopted the bistream network architecture to pursue a complementary status between RGB and depth, in which one of its streams is used for RGB saliency prediction and the other aims for depth saliency calculation. The final salient object detection results are obtained by feeding the previous RGB/depth saliency maps into a selective fusion module, and we exhibit the pipeline in Fig. 2A.

Corresponding author: Chenglizhao Chen, cclz123@163.com.

Fig. 1: Qualitative demonstrations towards the contribution of the depth (D) information. We use red boxes to highlight those false-alarm detections of the approach without using D (i.e., "w/o D"). Benefitting from D, those regions originally misdetected by the "w/o D" approach can now be detected correctly, and we use green boxes to highlight them.

However, there exists one major problem which seems to lead the classic bistream architecture based methods to reach a performance bottleneck: the depth quality usually differs from scene to scene, and the strong data adaptability of the current deep learning based techniques may lead the trained fusion subnet to bias towards the informative RGB component, producing mediocre detection even in the case that the depth channel is trustworthy; the quantitative results can be found in Fig. 2B.

To solve the aforementioned problem, we propose a channel-wise fusion scheme to integrate the depth channel into the RGB component, in which we use the depth (D) channel to cyclically replace each sub-channel of RGB, obtaining 3 independent 3-dimensional data, i.e., **D**GB, R**D**B and RG**D**, which can respectively be fed into any off-the-shelf RGB salient object detection model (e.g., [1]) to produce a much improved saliency estimation versus that solely based on the depth channel. The effectiveness of the proposed channel-wise fusion scheme is proven in our ablation experiments.

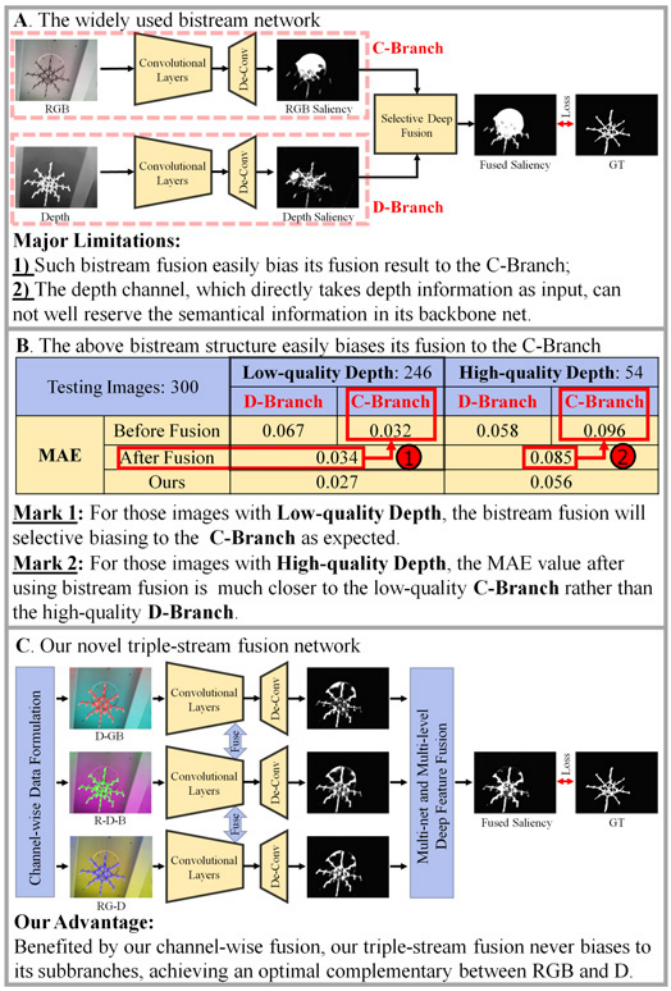

Fig. 2: The main highlight of this paper. Though the conventional bistream fusion will achieve better performance than either subbranch, its fused result easily biases to the RGB channels, failing to make full use of the depth information: for more information, refer to the quantitative results in subfigure B (more quantitative results can be found in the component evaluation section). Furthermore, we have listed the advantages of our newly proposed triple stream fusion network. All deep models mentioned are trained over an **identical** training set containing both high-quality and low-quality cases. We divide the testing set into the high-quality subset and the low-quality subset to highlight the biased fusion status of the conventional bistream networks.

In addition, to take full advantage of the novel data, we design our deep network as a triple-stream architecture, in which each subbranch will receive 1 of the 3 fused data as input to avoid the bias problem. Moreover, to ensure an optimal complementary status between RGB and D, we propose a lightweight recursive fusion strategy to interactively complement each subbranch of our network.

Additionally, we have conducted an extensive quantitative evaluation to validate the effectiveness of our method, in which we have compared our method with 16 SOTA methods over 5 widely adopted benchmark datasets. Overall, the main contributions of this paper can be summarized as follows:

- We have designed a novel data-level fusion scheme to integrate the RGB component with the depth channel, ensuring the high-quality and low-level saliency estimations;
- We have proposed a novel lightweight triple-stream fusion network to make full use of our newly formulated input data, ensuring an optimal complementary fusion status between RGB and D;
- We have conducted extensive validations and comparisons to show the effectiveness and advantages of our method, and our code[1] is also publicly available.

II. Related Works

In general, the depth-related saliency estimation is motivated by an assumption that the salient object should be located at a different depth layer from its nonsalient nearby surroundings. Thus, similar to the conventional RGB salient object detection methods, the key rationale for saliency estimation over the depth channel is conducting multi-level/multi-scale contrast computations.

A. Hand-Crafted Methods

As one of the most representative hand-crafted methods, Peng et al. [39] adopted the multi-level (e.g., local, global and background) contrast computation to obtain multi-contextual depth saliency. Similarly, Ren et al. [40] devised regional contrast to further boost the depth saliency quality. By using the off-the-shelf PageRank technique, the previously computed depth saliency features were fused with multiple RGB global priors as the final detection results. Although the regional contrast computation can result in robust detection performance for RGB-D images with simple backgrounds, it may occasionally produce massive false-alarm detections with cluttered backgrounds. To alleviate this problem, Feng et al. [41] proposed a region-wise angular contrast computation over depth information, in which the angular contrast degree, as an alternate saliency clue, may potentially be able to separate the salient object from its nonsalient nearby surroundings. Once the hand-crafted depth saliency was computed, these conventional methods would frequently follow the linear fusion scheme to integrate RGB saliency with depth saliency for an improved salient object detection result, in which the fusion weight can either be empirically assigned or adaptively formulated [42], [43].

B. Deep Learning based Methods

After entering the deep learning era, the SOTA methods have widely adopted the deep learning based techniques for high-performance RGB-D salient object detection. Qu et al. [33] utilized the compactness prior to guide the saliency detection over depth information, which would then be fused with its RGB saliency by using the CNNs based selective fusion module. However, the result may occasionally encounter various detection artifacts because both its RGB saliency and depth saliency were computed in a hand-crafted manner. To solve this issue, Shigematsu et al. [44] proposed a bistream

[1]Code&Data, https://github.com/XueHaoWang-Beijing/DRLF

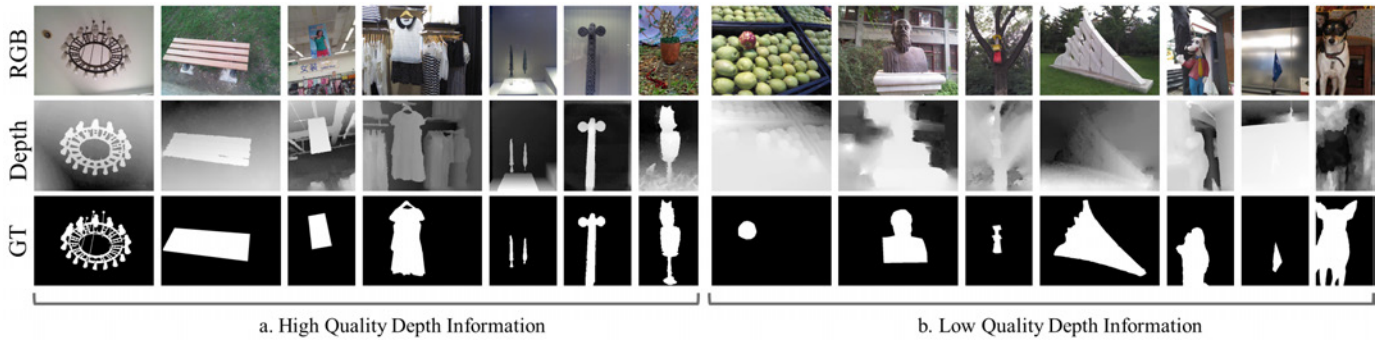

Fig. 3: Demonstration of the depth quality, in which the left component shows the high-quality cases, while the right component shows the low-quality cases. GT represents the human well-annotated binary saliency ground truth.

CNNs network to extract patch-wise deep features from RGB and depth, respectively. Then, both the RGB and depth-based deep features were concatenated and later fed into multiple fully connected layers to achieve a selective deep fusion.

Most recently, the FCNs network has been widely adopted to conduct end-to-end RGB-D salient object detection. Zhu et al. [45] followed the bistream network architecture, in which one of its subnets aimed to conduct FCNs based RGB saliency estimation, and the other focused on depth saliency revealing. Moreover, its subsequent RGB-D fusion module was also quite simple, which directly convolved the deep features generated by its depth and RGB subbranches to achieve an improved RGB salient object detection result. Although the depth features can greatly benefit the RGB saliency estimation in most cases, they may occasionally encounter failure detections when the depth information is less trustworthy (e.g., low-quality depth; see the right component of Fig. 3), leading to an even worse fused result. Thus, Wang et al. [38] weakly learned an additional fusion indicator, which was capable of adaptively ensuring its fusion procedure to bias towards its RGB component if the depth channel quality was predicted to be less trustworthy. Further, Zhao et al. [46] used an additional contrast loss to enhance the depth quality, achieving a much improved SOD result. Although many improvements have been made, the aforementioned methods still compute their depth saliency solely using depth information, which inevitably biases the fusion procedure towards the RGB component.

III. The Proposed Method

The semantic information usage is a vital factor to determine the overall SOD performance, and thus the deep learning based SOTA methods have widely adopted the off-the-shelf backbones (e.g., VGG16) to compute their high-discriminative deep features. Since most of these prevailing backbones are pre-trained using a large-scale "3-dimensional" training set with strong semantic knowledge (e.g., RGB images with human-labeled semantic categories), the deep features from these backbones are frequently embedded with strong semantic information even if these backbones are fine-tuned over other 3-dimensional SOD training sets.

In the face of 4-dimensional RGB-D data, we need to completely retrain these backbones if we seek to continue to take full advantage of the semantic information embedded in the pre-trained feature backbones. However, to the best of our knowledge, there exists no large-scale 4-dimensional RGB-D training set with strong semantic information. Thus, the existing RGB-D SOD approaches have considered the bistream network structure, including one RGB saliency branch, which takes 3-dimensional RGB data as input as usual, and one D saliency branch which duplicates the depth channel 2 times as "DDD". These individual saliency subbranches will later be selectively fused to output the final RGB-D saliency maps.

Though this methodology is reasonable and effective in most cases, it has reached a performance bottleneck because this methodology has overlooked one critical fact: the strong data adaptability of current deep learning based techniques may lead the selective fusion process to "extremely" bias towards one of its preceding subbranches if there exists a significant performance gap between these subbranches. To be more specific, as we have mentioned before, the RGB saliency subbranch usually significantly outperforms the D saliency subbranch, and, as a result, the valuable depth information, which is supposed to be capable of benefitting the SOD task, may be overwhelmed during the bistream selective deep fusion—a biased fusion process.

To overcome this issue, here we propose a simple yet effective way to fuse depth information with RGB information in channel-wise fashion, which is capable of making full use of both RGB and depth information and avoiding the aforementioned unbalanced fusion result.

A. Channel-Wise Data Fusion

The conventional methods compute their depth saliency using solely depth information, which easily leads to inferior depth saliency because the 3-dimensional "DDD" is still quite different from the "RGB" data: in fact, a large data gap exists, degenerating their feature backbones in providing semantic deep features.

To diminish such a "data gap", we advocate inserting the D channel into RGB channels cyclically: that is, we re-formulate the original 4-dimensional RGB-D data into 3 independent 3-dimensional data, i.e., **D**GB, R**D**B and RG**D**, in which we use the depth information to replace one channel of RGB data each time. Consequently, such new data are exactly within a 3-dimensional formulation, and, compared with the conventional

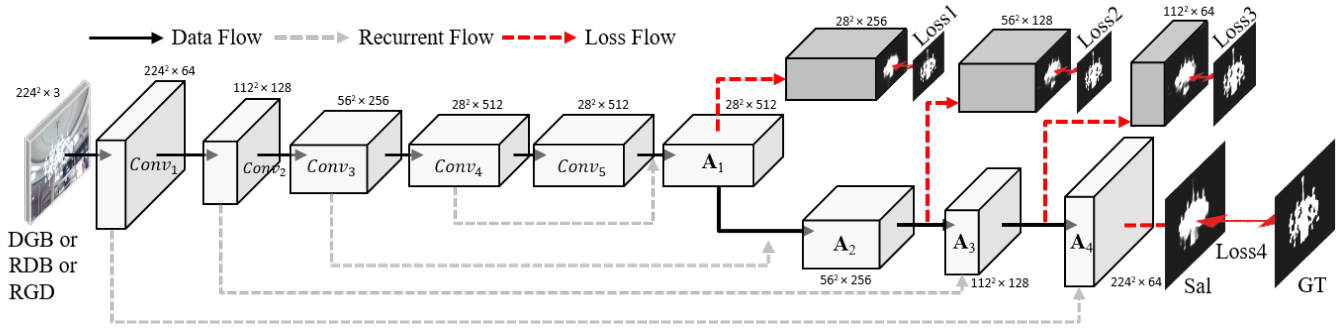

Fig. 4: Architecture of our backbone subnet.

"RGB+DDD", the feature backbones can output deep features with more semantic information after being fed by these new data.

B. Novel Backbone Network

So far, we can directly use the off-the-shelf pre-trained network (i.e., we simply choose the vanilla VGG16) to extract high-dimensional deep features respectively for each D-GB, R-D-B and RG-D, and we show the detailed network architecture in Fig. 4.

In our implementation, we have dropped the VGG16's last pool layer and the fully connected layers. The outputs of five convolutional blocks (separated by pool layers) are denoted as $\mathrm{Conv_i}, \mathrm{i} \in \{1, 2, 3, 4, 5\}$. In addition, we have modified the hyper-parameters of $\mathrm{Pool_4}$ to ensure abundant details in $\mathrm{Conv_5}$: i.e., the kernel size, padding number and stride in $\mathrm{Pool_4}$ layer are assigned to $\{3, 1, 1\}$, respectively.

To make full use of the multi-level deep features, we also recursively combine deep features from consecutive levels by feeding them into 2 convolutional layers with 1 up-sampling layer for skip connections. We denote the deep feature of each skip connection as $\mathbf{A}_i (i \in [1, 4])$, which is separately supervised by the ground truth to ensure the robustness of our network. Moreover, we normalize $\mathbf{A}_i$ before each skip connection. Thus, the outputs of multi-level skip connections can be formulated by Eq. 1.

$$\mathbf{A}_\mathrm{i} = \begin{cases} f\big(bn(\mathrm{Conv_j}) \circ bn(\mathbf{A}_{i-1})\big) & i \in [2, 4] \\ f\big(bn(\mathrm{Conv_4}) \circ bn(\mathrm{Conv_5})\big) & i = 1 \end{cases}, \quad (1)$$

where $\circ$ denotes the feature concatenation operation; $\mathbf{A}_i$ denotes the feature computed by the i-th skip connection; we set $j = 5 - i$, where bn denotes the batch normalization function; f denotes a 3×3 convolutional operation, which transforms input data into specific channel numbers as described in Fig. 4. Here we use Sal to represent the deepest data flow in our backbone subnet (see Eq. 2).

$$\mathrm{Sal} = f\Big(bn(\mathbf{A}_4) \circ bn\Big(f\Big(bn(\mathbf{A}_3) \circ bn\big(f\big(bn(\mathbf{A}_2) \circ bn(\mathbf{A}_1)\big)\big)\Big)\Big)\Big). \quad (2)$$

C. Multi-Level and Multi-Net Deep Feature Fusion

So far, we have obtained 3 independent branches by respectively feeding DBG, RDB, and RGD into 3 parallel backbone networks (Fig. 4). To achieve a complementary fusion status among these 3 subnets, here we propose a novel lightweight scheme to conduct multi-level and multi-net deep fusion (see the overall network architecture in Fig. 5).

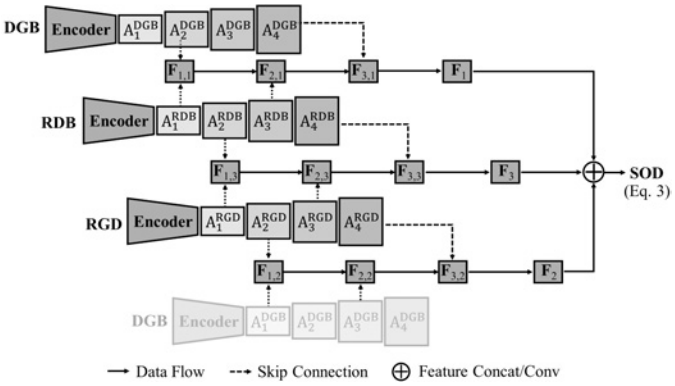

Fig. 5: The detailed network architecture of our triple-stream fusion network, in which the multi-level features ($\mathbf{A}_i, i \in \{1, 2, 3, 4\}$) are directly obtained from 3 parallel backbone branches (see details in Fig. 4). We use a hallucination of DGB (i.e., the bottom row) to facilitate a clear pipeline.

Since the 3-dimensional input data for each subbranch only consist of 2/3 color information (1/3 depth channel), we shall simultaneously resort multi-level deep features which are obtained from different branches to make full use of "all" color information. For example, we concatenate the deep feature $\mathbf{A}_1^{DGB}$ with the $\mathbf{A}_2^{RGD}$ and then convolve it as the fused $\mathbf{F}_{1,2}$. Then, we sequentially combine $\mathbf{F}_{1,2}$ with the deep feature $\mathbf{A}_3^{DGB}$ as the $\mathbf{F}_{2,2}$, and the $\mathbf{F}_{2,2}$ is further combined with the $\mathbf{A}_4^{RGD}$ as the $\mathbf{F}_{3,2}$, which is finally convoluted as the fused saliency $\mathbf{F}_2$.

Once all of these branch-wise saliency estimations have been obtained, we directly convolve $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$ as the final detection result (see Eq. 3).

$$\mathrm{SOD} = f(\mathbf{F}_1 \circ \mathbf{F}_2 \circ \mathbf{F}_3). \quad (3)$$

We have applied one loss function over each fusion procedure to ensure the correctness of the computed deep features. Thus, there are in total {1 (SOD) + 3 (backbones) * 4 (levels)

+ 3 (branches) * 3 (levels) = 22} loss functions in our triple-stream network, in which the total loss function can be formulated as Eq. 4.

$$\begin{aligned}\text{Loss} = \alpha_0 \cdot \text{L}(\text{SOD}) + \sum_{j=1}^{3}\sum_{i=1}^{4} \alpha_{i,j} \cdot \text{L}(\mathbf{A}_{i,j}) \\ + \sum_{r=1}^{3}\sum_{k=1}^{3} \alpha_{k,r} \cdot \text{L}(\mathbf{F}_{k,r}).\end{aligned} \tag{4}$$

Here, Loss denotes the total loss of our network, and $\text{L}(*)$ denotes the cross-entropy loss function; α_* denotes the weights of different loss functions, in which we empirically set $\alpha_{1,j}$, $\alpha_{2,j}$, $\alpha_{3,j}$, $\alpha_{4,j}$ and α_0 to {0.6, 0.7, 0.8, 0.9, 1}, respectively, and $\alpha_{*,r}$ are set to 0.9.

Specifically, each of these loss functions is proven to be indispensable in Sec. IV-H and Table VIII.

D. Network with Recurrent Saliency

The primary motivation of our method is to use the depth channel to complement RGB information for a high-performance salient object detection result. Our triple-stream network has already produced high-quality RGB-D saliency maps, which can be regarded as "corrected" depth maps to boost the overall SOD performance. We thus introduce a recurrent stage to ***refine*** our saliency maps. We use these saliency maps to replace the aforementioned depth part; i.e., we convert the original {**D**GB, R**D**B, RG**D**} into {**S**GB, R**S**B, RG**S**}, where **S** denotes the saliency prediction made by our triple-stream network (SOD, Eq. 3). Then, we recursively feed {SGB, RSB, RGS} into our triple-stream network again to further improve the detection performance (SOD^+, Eq. 5), and the quantitative results showing its effectiveness can be found in Table IV.

$$\text{SOD}^+ = \text{TriNet}\big(\text{R}, \text{G}, \text{B}, \text{TriNet}(\text{R}, \text{G}, \text{B}, \text{D})\big), \tag{5}$$

where, TriNet denotes our novel triple-stream network. Intuitively, we may continue this recurrence multiple times, but these new recurrent processes cannot achieve further performance improvement and may even lead to slight performance degeneration: this issue will be further investigated in the following ablation study (Sec. IV-J).

IV. Experiments

We shall firstly introduce all quantitative metrics and datasets (Sec. IV-A), and then we will provide 10 quantitative evaluations to validate the effectiveness of our method, including:

1) Sec. IV-B will verify whether the depth channel can really benefit the salient object detection task, providing solid evidence to readers regarding the effectiveness of the depth channel in boosting the overall detection performance.
2) Sec. IV-C will provide the detailed evidence to show the critical limitation of the conventional fusion approaches (i.e., they tend to bias towards the RGB subbranch) and verify the effectiveness of our method in handling this limitation.
3) Sec. IV-D will verify the effectiveness of the proposed data-level recombination scheme.
4) As a complementary part of Sec. IV-D, Sec. IV-F will further show the necessity of using the proposed 3-dimensional data-level recombination scheme, where we will provide several lines of quantitative evidence to demonstrate the advantages of our scheme with respect to the conventional schemes, e.g., the 4-dimensional RGB-D input.
5) Sec. IV-F will verify the effectiveness of the proposed multi-level and multi-net deep feature fusion strategy, in which we will compare it with the conventional linear fusion scheme.
6) Sec. IV-G will verify the effectiveness of the proposed lightweight fusion scheme, showing that we can achieve almost the same performance as the original dense fusion scheme if we use the proposed lightweight decoder.
7) Sec. IV-H will show the effectiveness of using multiple loss functions.
8) As a complementary part of Sec. IV-H, Sec. IV-I will provide a detailed ablation study towards different choices of loss weights.
9) Sec. IV-J will provide a detailed ablation study towards the effectiveness of the recurrent procedure mentioned in Sec. III-D.
10) Sec. IV-K will compare the proposed model with other SOTA models.

A. Datasets and Evaluation Metrics

We have quantitatively evaluated our method over 5 widely used benchmark datasets, e.g., NJUDS [47], NLPR [39], SSB [48], SSD [49] and DES [50]. A brief introduction will be given below:

NJUDS [47] contains 1985 pairs of color and depth images with manually labeled ground truth. Its color images are captured from the indoor environment, outdoor environment and stereo movies. The depth maps are generated by an optical flow method, and we normalized them to [0,255].

NLPR [39] is captured by Microsoft Kinect in both indoor and outdoor scenarios. It contains 1000 natural image pairs. Its depth maps are of high quality and contain sharp boundaries to distinguish the foreground and background. We normalized the depth images to [0,255]. Furthermore, we reversed the value of depth data so that the salient areas have high depth value.

SSB [48] is also named **STEREO** and contains 1000 RGB and depth image pairs. These images are captured from indoor places, outdoor nature scenes and stereo movies. However, numerous images in SSB are repeated in NJUDS and NLPR datasets.

SSD [49] is selected from the stereo movies and consists of 80 indoor/outdoor stereo images. Most scenes in this dataset contain multiple salient objects. Specifically, the human-annotated ground truths in this dataset are quite different from the other datasets. In this dataset, scenes clearly contain multiple salient objects, while the saliency ground truth annotations of these images tend to regard only one of them as the salient one. The details are shown in Fig. 7. Thus, this dataset is more challenging than others.

DES [50] is captured by Microsoft Kinect and contains 135 indoor stereo images. However, due to the limited quality, the depth channels of most RGB-D images in this dataset may only be able to coarsely locate the salient object.

Evaluation Metrics We have evaluated the performance of our method and other SOTA methods using 8 widely adopted metrics [51], e.g., S-measure [52], E-measure [53] (adaptive, mean, max), F-measure (adaptive, mean, max) and mean absolute error (MAE).

B. The Effectiveness of Depth Channel

To verify the advantage of using depth information, we have conducted the quantitative comparisons, i.e., the proposed network ***without using*** depth information vs. the proposed network ***using*** depth information. As shown in Table I, the bottom row has achieved the best performance, using the depth information, while the first two rows have exhibited inferior performances. Since all of these results are obtained via an identical network structure, the performance margins are mainly induced by the depth information, exhibiting the advantages of using depth information.

TABLE I: Quantitative comparisons between models without using depth information (i.e., the first two rows) and the model using depth information (i.e., the bottom row). Specifically, for the 2-dimensional input case (i.e., the "GB+RB+RG" input), we use a zero matrix as the additional dummy channel to meet the 3-dimensional input requirement of the proposed network.

Channel Information	NJUDS			NLPR			DES		
	meanF ↑	maxF ↑	MAE ↓	meanF ↑	maxF ↑	MAE ↓	meanF ↑	maxF ↑	MAE ↓
RGB+RGB+RGB	.822	.848	.067	.835	.865	.037	.759	.788	.046
GB+RB+RG	.821	.852	.070	.832	.861	.038	.759	.799	.047
DGB+RDB+RGD	**.853**	**.881**	**.058**	**.844**	**.872**	**.035**	**.846**	**.869**	**.031**

Moreover, from the perspective of RGB-D salient object detection, the rationale of using depth information lies in providing an additional "venue" for networks to separate/highlight salient objects from their nonsalient nearby surroundings, aiming for improving the overall detection performance in terms of detection accuracy and completeness. In this regard, any other "venues" which may potentially be capable of benefitting the salient object detection task can be used to replace the depth channel. However, the problem is how to obtain such unnatural venues. Indeed, the saliency maps produced by other state-of-the-art (SOTA) approaches apparently belong to such venues, but, compared with the depth information, these saliency masks may not be good choices because the saliency maps are originally revealed from the RGB information, and we may become self-trapped if we use them again to complement the RGB channel.

In sharp contrast, the depth information—a typical natural source to complement RGB information—can be used to sense various low-level saliency cues because the depth layers of salient objects usually exhibit large differences versus the nonsalient nearby surroundings.

To validate the above claim, we have conducted the ablation study to show the effectiveness of depth information. As shown in Table II, with the help of depth information, our method (i.e., RGB+SOD) achieves up to 8.2% performance improvement over other RGB-masks based approaches (i.e., RGB+$\mathrm{M}_{1/2/3}$), showing that depth information is able to offer additional cues to improve the performance of salient object detection tasks. Furthermore, the overall performance of using saliency masks (e.g., RGB+SOD and RGB+$\mathrm{M}_{1/2/3}$) to replace depth information is heavily dependent on the quality of the saliency mask. A high-quality saliency mask will improve the overall performance (e.g., the RGB+SOD), while a low-quality saliency mask may even result in decreased overall performance (e.g., RGB+M_1). On the other hand, performance improvement enabled by saliency masks is clearly limited, which is quite marginal versus that of the depth information, e.g., $\{$RGB+D$\}$ vs. $\{$RGB+SOD$\}$.

TABLE II: Ablation experiments regarding different combinations, where RGB+$\mathrm{M}_{1/2/3}$ respectively denote 3 different approaches which use saliency masks (i.e., M_1 [54], M_2 [55] and M_3 [56]) to replace the depth information. "RGB+SOD" denotes the results of our approach.

D/Mask	NJUDS			NLPR			DES		
	meanF ↑	maxF ↑	MAE ↓	meanF ↑	maxF ↑	MAE ↓	meanF ↑	maxF ↑	MAE ↓
RGB+M_1	.837	.861	.061	.837	.865	.035	.774	.803	.041
RGB+M_2	.846	.870	.058	.853	**.882**	.034	.787	.835	.041
RGB+M_3	.838	.863	.062	.846	.873	.034	.789	.823	.040
RGB+SOD	**.858**	**.883**	**.055**	**.854**	.880	**.032**	.832	**.869**	**.030**
RGB+D	.853	.881	.058	.844	.872	.035	**.846**	.869	.031

C. Bistream Fusion Biasing Quantitative Results

The biasing tendency in the conventional bistream fusion networks (see the first two rows in Table III) is mainly induced by the unbalanced status between their two subbranches. As shown in the "high-quality depth" columns, we can see that there are in total 300 images in the NLPR testing set, and only 54 depth maps are capable of producing more distinguishing salient features than RGB images. The strong data adaptability of the current deep learning based techniques leads the bistream fusion subnet to bias towards the informative RGB component (MAE is 0.096), producing a mediocre detection (MAE is 0.085) even in the case that the depth channel is trustworthy (MAE is 0.058). On the other hand, we can see from the 4th to 7th rows that most of the SOTA methods (including bistream PCA [57], MMCI [58], AFNet [38] and triple-stream TANet [59]) bias towards the color information even if the depth information is more reliable in some cases. Benefitting from the single-stream structure, we can see that CPFP [46] (the eighth row in Table III) decreases the tendency of biasing, but does not solve the problem. Compared with the above mentioned SOTA methods, our model overcomes this weakness by balancing its subbranches via balanced input data, achieving a much improved performance in cases with either high-quality (up to 37% improvement in maxF) or low-quality (up to 47% improvement in maxF) depth information (see the third row in Table III).

D. Effectiveness of Data-level Fusion

We design conventional single-stream/bistream/triple-stream networks based on the VGG16 network, where the fusion strategies of the bistream/triple-stream networks are either linear concatenation (denoted by LC) or the proposed multi-level and multi-net deep feature fusion model (denoted by MF). To validate the effectiveness of the proposed

TABLE III: Bistream fusion biasing quantitative results, in which "SIN","BI" and "TRI" are the abbreviations of "single-stream", "bistream" and "triple-stream", respectively. In general, both the conventional bistream fusion and the single/multiple stream SOTA methods will bias to the color information regardless of the quality of depth maps. In sharp contrast, our method can handle this problem well. All deep models mentioned are trained over an **identical** training set containing both high-quality and low-quality cases. We divide the testing set into the "high-quality" subset and the "low-quality" subset to highlight the biased fusion status of the conventional bistream networks.

NLPR [39] (300)		Low-quality Depth: 246		High-quality Depth: 54	
		D-Branch	C-Branch	D-Branch	C-Branch
MAE	Before Fusion	.067	.032	**.058**	.096
	After Fusion	.034		.085	
	Ours/TRI	.027		**.056**	
	TANet/TRI [59]	.033		.073	
	PCA/BI [57]	.036		.080	
	MMCI/BI [58]	.053		.088	
	AFNet/BI [38]	.052		.089	
	CPFP/SIN [46]	.028		.068	
DES [50] (135)		Low-quality Depth: 83		High-quality Depth: 52	
		D-Branch	C-Branch	D-Branch	C-Branch
MAE	Before Fusion	.042	.024	**.044**	.084
	After Fusion	.025		.072	
	Ours/TRI	.018		**.049**	
	TANet/TRI [59]	.030		.071	
	PCA/BI [57]	.030		.079	
	MMCI/BI [58]	.050		.089	
	AFNet/BI [38]	.053		.092	
	CPFP/SIN [46]	.021		.064	

data-level fusion scheme, we respectively feed different input data into the single-stream network. As shown in Table IV, our data-level fusion schemes, e.g., DGB, RDB and RGD, achieve obvious performance improvements versus the original RGB saliency maps (1.9% in maxF) and depth saliency maps (14.8% in maxF). Additionally, we may easily notice that the single-stream network using our novel data-level fusion scheme has achieved comparable performance to the bistream network using the conventional RGB-D input data. In addition, as shown in the middle rows of Table IV, the conventional bistream network using our data-level fused input can achieve significant performance improvements: e.g., the {RDB+RGD} improves the conventional {RGB+D} by almost 1% in maxF. Actually, we believe that the above results are mainly induced by the following three aspects:

1) The conventional bistream network, which takes the depth channel as the sole input for its depth branch, cannot make full use of its backbone network, failing to obtain deep features with meaningful semantic information;

2) The conventional bistream network easily biases its fusion procedure to the color branch due to the inferior performance of its depth branch;

3) Our data-level fusion is simple yet effective with respect to complementing the color information with the depth information. Moreover, it can effectively adapt to the backbone network because of the 3-dimensional data structure, obtaining useful semantic information for saliency detection.

TABLE IV: Component evaluation results over NLPR dataset. "RGB", "D", "DGB", "RDB" and "RGD" denote different input data combinations of the single-stream/bistream/triple-stream network. "LC" and "MF" represent the different fusion strategies, e.g., linear concatenation and the proposed multi-level and multi-net deep feature fusion model. ↑ means that the larger one is better, and ↓ denotes that the smaller one is better.

	Combinations	meanF ↑	maxF ↑	MAE ↓
Single-stream	{RGB}	.805	.837	.051
	{D}	.712	.743	.115
	{RGB-D}	.772	.826	.053
	{DGB}	.821	.853	.045
	{RDB}	.819	.853	.043
	{RGD}	.821	.852	.042
Bistream	{RGB+D}	.825	.853	.042
	{DGB+RDB}	.827	.858	.040
	{RDB+RGD}	.831	.862	.039
	{RGD+DGB}	.832	.861	.039
Tri-Stream	LC{RGB+D+D}	.822	.851	.041
	LC{RGB+RGB+D}	.829	.860	.040
	LC{DGB+RDB+RGD}	.838	.862	.038
	MF{RGB+D+D}	.828	.858	.040
	MF{GB+RB+RG}	.832	.861	.038
	MF{RGB+RGB+D}	.834	.865	.038
	MF{RGB+RGB+RGB}	.835	.865	.037
	MF{DGB+RDB+RGD}	.844	.872	.035
	Final (Recurrent)	.854	.880	.032

On the other hand, each of the 3-dimensional data outperforms the 4-dimensional RGB-D data in the single-stream network, even though the 4-dimensional data include the information of one more channel. The main reason for this is that the widely used feature backbones are pre-trained using the 3-dimensional training set with strong semantic information, while such pre-learned semantic information may be lost if we fine-tune it over 4-dimensional data (i.e., RGB-D), exhibiting another advantage of our data-wise fusion.

E. Reasoning behind Channel-Wise Data Fusion

It is well known that the performance of feature backbones usually plays an important role in determining the overall detection performance, and the feature backbones used in the salient object detection field are usually trained over other large-scale training sets (e.g., the ImageNet). Since the function of the depth channel is quite similar to RGB channels, providing some possible cues to separate different objects, the feature gap between the original RGB and the newly re-formulated data ({D+RG}, {D+RB}, and {D+GB}) shall be marginal. Thus, by using such re-formulated data, the pre-trained feature-backbones are still capable of providing meaningful and discriminative deep features even after being

fine-tuned using other training sets, e.g., the widely used RGB-D training set with 2050 images. However, there usually exist large differences between the 3-dimensional RGB and the 2-dimensional R+D (or G/B+D), which make the pre-trained feature backbones unsuitable for the 2-dimensional input data, requiring a complete new training. Unfortunately, the widely used RGB-D training set only consists of 2050 training instances, which are clearly insufficient to train a complete new feature backbone. Thus, due to the above mentioned aspects, we decide to resort to the 3-dimensional data formulations, i.e., {D+RG}, {D+RB}, and {D+GB}.

Specifically, because the off-the-shelf feature backbones are all trained using the RGB training set, significant performance degeneration will result if we choose to use the luminance & chrominance formulation (i.e., the YUV color space). To make the above explanations more convincing, we have tested the performance of the above mentioned 2-dimensional case and the YUV case. As shown in Table V, the second row has achieved the best performance, while other cases cannot perform well because of the large gap between these data formulations and the original RGB training space.

TABLE V: Ablation experiments of different data formulations. We highlight the best results with bold typeface.

Color Space	NJUDS			NLPR			DES		
RGB/YUV	meanF	maxF	MAE	meanF	maxF	MAE	meanF	maxF	MAE
RD+GD+BD	.841	.875	.062	.840	.870	.038	.817	.848	.036
RGD+RDB+DGB	**.853**	**.881**	**.058**	**.844**	**.872**	**.035**	**.846**	**.869**	**.031**
YUD+YDV+DUV	.741	.756	.100	.740	.767	.055	.646	.775	.063

F. Effectiveness of the Proposed Triple-Stream Fusion Strategy

We also explore the effectiveness of the proposed multi-level and multi-net deep feature fusion strategy. As shown in Table IV, Tri-Stream, the proposed fusion strategy, achieves a significant performance improvement; i.e., the proposed fusion strategy (denoted by MF) outperforms the conventional linear concatenation (denoted by LC) using identical input data (0.8%/0.6%/1.2% in maxF over {RGB+D+D}, {RGB+RGB+D} and {DGB+RDB+RGD}, respectively). Compared to the conventional fusion scheme, the proposed multi-level and multi-net feature fusion scheme offers two major advantages:

1) It achieves an optimal complementary fusion status among its triple-stream subbranches;

2) It fuses the multi-scale deep features of different subbranches, and thus, its final prediction can effectively preserve tiny details.

G. Lightweight Fusion Strategy

To validate the lightweight structure of the proposed fusion strategy, we 1) compare the number of parameters with other SOTA methods and 2) compare the performance over evaluation metrics with dense fusion schemes. As shown in Table VI, we have listed the parameter quantities in the backbone and fusion subnetwork individually to show the computational expenses. All compared methods are based on the VGG network. The unit of the number of parameters is *"million"*. At first glance, the proposed method has a larger number of parameters than single-stream CPFP [46], i.e., the number of parameters in CPFP is approximately 15 (one backbone) + 57.9 (fusion) = 72.9 in total, yet the number of parameters in our model is 3 × 15 (three backbones) + 46.8 (fusion) = 91.8 in total. Since our method has adopted the triple-stream network architecture which consists of 3 backbones, and these backbones constitute almost 49% parameters and the fusion connections constitute the rest 51%, this is a much lower ratio than the CPFP method (79%). Thus, from the perspective of a triple-stream network, our model is lightweight-designed and the fusion connections of our method are very sparse (with only 12 short-connections shown in Fig. 5). Our model achieves a speed of 15.6 FPS, which is faster than most of the STOA methods.

TABLE VI: Comparisons regarding the number of parameters between our model and other SOTA methods (unit: million). From the perspective of fusion-total ratio, our fusion scheme achieves strong inter-subnet interactions, yet at the lowest computational cost.

Method	Total	Backbone	Fusion	F-T Ratio	FPS
CPFP [46]	72.90	1 × 15	57.90	79%	9.2
TESF [35]	98.30	1 × 15	83.30	85%	9.8
PCA [57]	138.7	2 × 15	108.7	78%	15.3
DCFF [60]	244.8	2 × 15	214.8	88%	13.1
MMCI [58]	241.7	2 × 15	211.7	88%	19.5
TANet [59]	247.5	3 × 15	202.5	82%	14.2
Our Model	91.80	3 × 15	46.80	51%	15.6

Furthermore, we compare our method with other triple-stream fusion schemes, which are shown in Fig. 6. Notice that we only take the fusion procedures between **D**GB and RG**D** branches for example. Since the 3-dimensional input data for each subbranch only consist of 2 color channels and 1 depth channel, we sparsely combine {DGB, RDB}, {RDB, RGD} and {RGD, DGB} branches first, and then iteratively integrate the outputs of these three subbranches to obtain our final detection.

As shown in Fig. 6A and B, we evaluate the performance of the fusion scheme without/with use of the multi-scale information. Compared with our fusion scheme, Fig. 6C shows a relatively more dense manner to fuse multi-scale and multi-net features. The performances of different fusion schemes are respectively reported in Table VII. In comparing the fusion manners C and D (i.e., the last two rows in Table VII), it is obvious that approximately 0.4% performance improvement may be achieved if we take all multi-scale and multi-net features into consideration. We can observe that our sparse fusion scheme (Fig. 6D) performs comparably to the dense fusion scheme (Fig. 6C), saving almost 37% of parameters in total because our proposed fusion strategy considers all channel information and different scale features by a crossing connection manner. Thus, as one of the key contributions, our fusion scheme achieves strong inter-subnet interactions, yet at the lowest computational cost, and this is why we claim our method as a lightweight design.

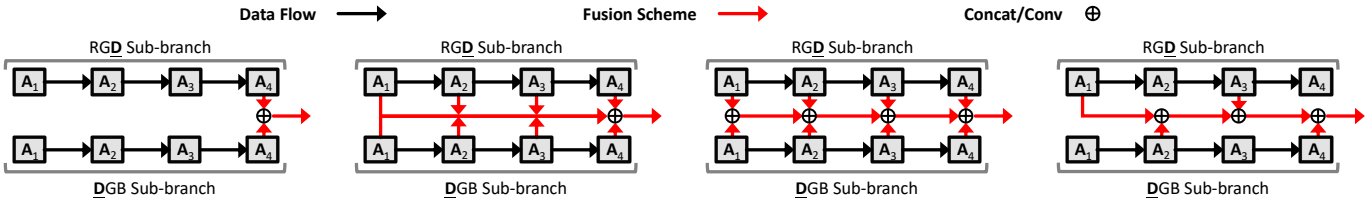

Fig. 6: Comparisons among different fusion schemes, in which $\mathbf{A}_i, i \in \{1, 2, 3, 4\}$ represent multi-level features of our backbones, which are parallel and identical to those in Fig. 5. Our fusion scheme is shown as subfigure D, which is lightweight-designed and performs well: please refer to the corresponding quantitative results in Table VII.

TABLE VII: Ablation study over triple-stream fusion schemes. The detailed structures of different fusion schemes are shown in Fig. 6. The unit of the number of parameters is "million". We highlight the best performances with bold typeface.

Fig. 6	Para-meters	NJUDS			NLPR			DES		
		meanF	maxF	MAE	meanF	maxF	MAE	meanF	maxF	MAE
A	86.2	.841	.865	.063	.838	.863	.038	.836	.859	.039
B	90.4	.849	.876	.061	.841	.868	.037	.840	.864	.034
C	144.9	**.855**	**.882**	**.057**	**.845**	.871	**.035**	**.846**	.868	**.031**
D	91.8	.853	.881	.058	.844	**.872**	**.035**	**.846**	**.869**	**.031**

H. Effectiveness of Multiple Loss Functions

There are 22 loss functions in our model: though numerous in quantity, all of them are indispensable. The usage of such a large amount of loss functions is inspired by the classic DSS [1], in which the single-stream method has adopted 6 loss functions in total to ensure high-quality side-outputs. The DSS has proven that discriminative features in different hidden layers will complement each other, and it is required to assign an individual loss function for each hidden layer to obtain such features.

Many existing models have followed the DSS, but they fail to explore the performance improvement by using multiple loss functions in different stages. To clearly distinguish these loss functions (Eq. 4), we divide them into three parts according to their supervised stages: i.e., we refer to the loss functions used in the backbone/branch-fusion/final-result as "$\mathrm{L_A}$ (i.e., $\mathrm{L}(\mathbf{A})$ in Eq. 4)", "$\mathrm{L_F}$ (i.e., $\mathrm{L}(\mathbf{F})$ in Eq. 4)", and "$\mathrm{L_S}$ (i.e., $\mathrm{L}(\mathrm{SOD})$ in Eq. 4)", respectively. We then conduct an extensive ablation study to validate the effectiveness of these loss functions.

As shown in Table VIII, we regard the $\mathrm{L_S}$ as the baseline, and other loss functions will be respectively applied to it. With the increasing number of loss functions, the network achieves steady performance improvement (i.e., 6.2% in maxF, which can be seen in the first four rows in Table VIII). Compared with the loss functions in the branch fusion stage (marked as $\mathrm{L_{S+F}}$), the loss functions in the backbone stage (marked as $\mathrm{L_{S+A}}$) have produced larger performance improvement (1.7% in maxF).

I. Choice of Loss Weights

The hyper-parameters of our approach mainly refer to the loss weights mentioned in Eq. 4, including $\alpha_{1,j}$, $\alpha_{2,j}$, $\alpha_{3,j}$, $\alpha_{4,j}$, α_0 and $\alpha_{*,r}$, which are empirically assigned to {0.6, 0.7, 0.8, 0.9, 1, 0.9}, respectively. This implementation is based on the fact that features from shallower layers of U-Net usually retain more information than those from the deep layers, and thus we assign large weights to the loss functions of the shallower layers. We set the loss weights (e.g., $\alpha_{*,r}$) of the fused layers equally to 0.9 because these layers usually contribute to the final output in a progressive manner. Specifically, to enhance the importance of the last layer that combines all branch-wise saliency as the final saliency, we set its loss weight to 1.

TABLE VIII: Ablation study over different combinations of loss functions. We divide the loss functions into three parts (Losses A, F, S) according to their applied stages. The loss weights are assigned as Eq. 4. $\mathrm{L^*_{S+F+A}}$ denotes that we equally set all hyper-parameters in the total loss to 1. The best performances are highlighted with bold typeface.

Loss Functions	NJUDS			NLPR			DES		
	meanF	maxF	MAE	meanF	maxF	MAE	meanF	maxF	MAE
$\mathrm{L_S}$	.813	.842	.077	.79	.821	.049	.767	.826	.049
$\mathrm{L_{S+A}}$	.852	.875	.061	.832	.858	.038	.815	.842	.036
$\mathrm{L_{S+F}}$	.848	.867	.062	.818	.842	.040	.790	.828	.039
$\mathrm{L_{S+F+A}}$	**.853**	**.881**	**.058**	**.844**	**.872**	**.035**	**.846**	**.869**	**.031**
$\mathrm{L^*_{S+F+A}}$	.843	.868	.062	.810	.835	.041	.788	.823	.040

In order to validate the effectiveness of our empirical assignment for these hyper-parameters, we have conducted another verification, in which we set all hyper-parameters to 1 (marked as $\mathrm{L^*_{S+F+A}}$ in the 5th row of Table VIII). Obviously, our method is slightly sensitive to these hyper-parameters, retaining its overall performance within approximately 4.4%.

TABLE IX: Ablation study over different loss weights. ↑ means that the larger one is better, and ↓ denotes that the smaller one is better. The best results are highlighted with bold typeface.

Hyper-parameters						NJUDS			NLPR			DES		
$\alpha_{1,j}$	$\alpha_{2,j}$	$\alpha_{3,j}$	$\alpha_{4,j}$	α_0	$\alpha_{*,r}$	meanF ↑	maxF ↑	MAE ↓	meanF ↑	maxF ↑	MAE ↓	meanF ↑	maxF ↑	MAE ↓
0.6	0.7	0.9	0.9	1	0.9	**.853**	**.881**	**.058**	**.844**	**.872**	**.035**	**.846**	**.869**	**.031**
0.6	0.7	0.9	0.9	1	1	.852	.879	.059	.841	.872	.036	.835	.864	.033
0.2	0.4	0.6	0.8	1	0.9	.851	.878	.059	.842	.870	.036	.828	.852	.034
0.2	0.4	0.6	0.8	1	1	.852	.877	.060	.840	.871	.037	.830	.851	.035
1	1	1	1	1	1	.843	.868	.062	.810	.869	.041	.788	.823	.040

In addition, we have provided the ablation study towards the choices of these hyper-parameters, and the quantitative results can be found in Table IX. The first two rows indicate that we should set the largest weight to α_0, and the following 2 rows show that we should not assign excessively small values to the

TABLE X: Quantitative comparison results in terms of S-measure, E-measure, MAE and F-measure over 5 challenging benchmark datasets. ↑ denotes that larger is better, and ↓ denotes that smaller is better. The best results are highlighted with bold typeface. The advantage of our method over the SSD dataset (still belonging to the top-3 methods) is not obvious due to the reason discussed in section IV-K.

	Metric	LHM 2014	ACSD 2014	GP 2015	LBE 2016	DCMC 2016	SE 2016	CDCP 2017	MDSF 2017	DF 2017	CDB 2018	CTMF 2018	PCA 2018	AFNet 2019	MMCI 2019	TANet 2019	CPFP 2019	Ours SOD	Ours SOD$^+$
NJUDS [47]	**Sm** ↑	.514	.699	.527	.695	.686	.664	.669	.748	.763	.624	.849	.877	.772	.858	.878	.878	.886	**.886**
	adpE ↑	.708	.786	.716	.791	.791	.772	.747	.812	.835	.745	.864	.896	.846	.878	.893	.895	.899	**.901**
	meanE ↑	.447	.593	.466	.655	.619	.624	.706	.677	.696	.565	.846	.895	.826	.851	.895	**.910**	.901	.909
	maxE ↑	.724	.803	.703	.803	.799	.813	.741	.838	.864	.742	.913	.924	.853	.915	.925	.923	.926	**.926**
	adpF ↑	.638	.696	.655	.740	.717	.734	.624	.757	.784	.648	.788	.844	.768	.812	.844	.837	.843	**.849**
	meanF ↑	.328	.512	.357	.606	.556	.583	.595	.628	.650	.482	.779	.840	.764	.793	.841	.850	.853	**.858**
	maxF ↑	.632	.711	.647	.748	.715	.748	.621	.775	.804	.648	.845	.872	.775	.852	.874	.877	.881	**.883**
	MAE ↓	.205	.202	.211	.153	.172	.169	.180	.157	.141	.203	.085	.059	.100	.079	.060	**.053**	.058	.055
STERE [48]	**Sm** ↑	.562	.692	.588	.660	.731	.708	.713	.728	.757	.615	.848	.875	.825	.873	.871	.879	.883	**.888**
	adpE ↑	.770	.793	.784	.749	.831	.825	.796	.830	.838	.808	.864	.897	.886	.901	.906	.903	.911	**.915**
	meanE ↑	.484	.592	.509	.601	.655	.665	.751	.614	.691	.561	.841	.887	.872	.873	.893	**.912**	.898	.911
	maxE ↑	.771	.806	.743	.787	.819	.846	.786	.809	.847	.823	.912	.925	.887	.927	.923	.925	.924	**.929**
	adpF ↑	.703	.661	.711	.595	.742	.748	.666	.744	.742	.713	.771	.826	.807	.829	.835	.830	.837	**.845**
	meanF ↑	.378	.478	.405	.501	.590	.610	.638	.527	.617	.489	.758	.818	.806	.813	.828	.841	.838	**.850**
	maxF ↑	.683	.669	.671	.633	.740	.755	.664	.719	.757	.717	.831	.860	.823	.863	.861	.874	.871	**.878**
	MAE ↓	.172	.200	.182	.250	.148	.143	.149	.176	.141	.166	.086	.064	.075	.068	.060	.051	.055	**.050**
DES [50]	**Sm** ↑	.578	.728	.636	.703	.707	.741	.709	.741	.752	.645	.863	.842	.770	.848	.858	.872	**.896**	.895
	adpE ↑	.761	.855	.785	.911	.849	.852	.816	.869	.877	.868	.911	.912	.874	.904	.919	.927	**.958**	.954
	meanE ↑	.477	.612	.503	.649	.632	.707	.748	.621	.684	.572	.826	.838	.810	.825	.863	.889	.902	**.910**
	maxE ↑	.653	.850	.670	.890	.773	.856	.811	.851	.870	.830	.932	.893	.881	.928	.910	.923	**.947**	.940
	adpF ↑	.631	.717	.686	.796	.702	.726	.625	.744	.753	.729	.778	.782	.730	.762	.795	.829	**.874**	.868
	meanF ↑	.345	.513	.412	.576	.542	.617	.585	.523	.604	.502	.756	.765	.713	.735	.790	.824	.846	**.852**
	maxF ↑	.511	.756	.597	.788	.666	.741	.631	.746	.766	.723	.844	.804	.729	.822	.827	.846	.869	**.869**
	MAE ↓	.114	.169	.168	.208	.111	.090	.115	.122	.093	.100	.055	.049	.068	.065	.046	.038	.031	**.030**
NLPR [39]	**Sm** ↑	.630	.673	.654	.762	.724	.756	.727	.805	.802	.629	.860	.874	.799	.856	.886	.888	.899	**.903**
	adpE ↑	.813	.742	.804	.855	.786	.839	.800	.812	.868	.809	.869	.916	.884	.872	.916	.924	.934	**.936**
	meanE ↑	.560	.578	.571	.719	.684	.742	.781	.745	.755	.565	.840	.887	.851	.841	.902	.918	.913	**.922**
	maxE ↑	.766	.780	.723	.855	.793	.847	.820	.885	.880	.791	.929	.925	.879	.913	**.941**	.932	.937	.939
	adpF ↑	.664	.535	.659	.736	.614	.692	.608	.665	.744	.613	.724	.795	.747	.730	.796	.823	.839	**.843**
	meanF ↑	.427	.429	.451	.736	.543	.624	.609	.649	.664	.422	.740	.802	.755	.737	.819	.840	.844	**.854**
	maxF ↑	.622	.607	.611	.745	.648	.713	.645	.793	.778	.618	.825	.841	.771	.815	.863	.867	.872	**.880**
	MAE ↓	.108	.179	.146	.081	.117	.091	.112	.095	.085	.114	.056	.044	.058	.059	.041	.036	.035	**.032**
SSD [49]	**Sm** ↑	.566	.675	.615	.621	.704	.675	.603	.673	.747	.562	.776	**.841**	.714	.813	.839	.807	.836	.835
	adpE ↑	.730	.765	.795	.729	.786	.778	.705	.772	.812	.737	.838	**.886**	.803	.860	.879	.832	.878	.880
	meanE ↑	.498	.566	.529	.574	.646	.631	.676	.576	.690	.477	.796	.856	.762	.796	**.861**	.839	.847	.853
	maxE ↑	.717	.785	.782	.736	.786	.800	.700	.779	.828	.698	.865	.894	.807	.882	**.897**	.852	.870	.870
	adpF ↑	.580	.656	.749	.613	.679	.693	.522	.674	.724	.628	.710	.791	.694	.748	.767	.726	.801	**.801**
	meanF ↑	.367	.469	.453	.489	.572	.564	.515	.470	.624	.347	.689	.777	.672	.721	.773	.747	.786	**.791**
	maxF ↑	.568	.682	.740	.619	.711	.710	.535	.703	.735	.592	.729	.807	.687	.781	.810	.766	**.812**	.810
	MAE ↓	.195	.203	.180	.278	.169	.165	.214	.192	.142	.196	.099	**.062**	.118	.082	.063	.082	.068	.066

loss weights in shallower layers. The bottom row illustrates that we may reach a clear sub-optimal performance if we assign all loss functions equally to 1.

J. Ablation Study on the Recurrent Times

As we have mentioned in Sec. III-D, the overall performance of our method can be further improved by using the recurrent procedure, i.e., replacing the depth channel of the input image with the corresponding saliency map estimated by the already trained network and performing another round of network training. Intuitively, this recurrent process of injecting the output as the input can be performed multiple times: however, based on our experience, only the first round of the recurrent procedure can persistently manifest performance gain, while the subsequent ones may even degenerate the overall performance. To verify this issue, we have conducted an additional quantitative evaluation, and the results can be found in Table XI, where "SOD$^{+\mathrm{n}}$" denotes the n^{th} round of recurrent procedure. Though some marginal improvements over some metrics can be observed, we suggest using the recurrent procedure only once to strike the optimal trade-off between performance and efficiency.

TABLE XI: Comparisons regarding the recurrent stages. "SOD" denotes our initial results, "SOD$^{+\mathrm{n}}$" denotes the results applying the recurrent procedure n times. We highlight the best results with bold typeface.

Model	NJUDS			NLPR			DES			STERE			SSD		
	meanF	maxF	MAE	meanF	maxF	MAE	meanF	maxF	MAE	meanF	maxF	MAE	meanF	maxF	MAE
SOD	.853	.881	.058	.844	.872	.035	.846	**.869**	.031	.838	.871	.055	.786	**.812**	.068
SOD^{+1}	.858	**.883**	.055	.854	.880	.032	**.852**	**.869**	**.030**	.850	.878	.050	**.791**	.810	**.066**
SOD^{+2}	.859	.879	.054	.859	**.883**	.031	.850	.865	**.030**	.855	.879	.049	.790	.807	**.066**
SOD^{+3}	**.860**	.878	**.053**	**.861**	**.883**	**.030**	**.852**	.865	**.030**	.858	**.880**	.048	.789	.803	**.066**
SOD^{+4}	**.860**	.877	**.053**	.859	.882	**.030**	.850	.864	**.030**	.859	**.880**	**.047**	.785	.801	**.066**
SOD^{+5}	**.860**	.876	**.053**	.859	.880	**.030**	.850	.863	**.030**	**.860**	.879	**.047**	.785	.800	**.066**

K. Performance Comparisons

We have compared our method with 16 SOTA RGB-D salient object detection methods over all of the adopted 5 benchmark datasets. The adopted SOTA methods include LHM [39], ACSD [61], GP [40], LBE [41], DCMC [42], SE [62],

Fig. 7: Demonstration of incomplete human-annotated ground truth masks (GT) in SSD, which are mainly induced by subjective annotations.

CDCP [63], MDSF [64], DF [33], CDB [65], CTMF [66], P-CA [57], AFNet [38], MMCI [58], TANet [59] and CPFP [46]. For a fair comparison, the saliency maps/executable codes of the compared methods are all provided by the authors with parameters/implementations unchanged.

As shown in Table X, our method can persistently outperform all compared SOTA methods in terms of 8 metrics over all 5 adopted benchmark datasets. Specifically, our method has achieved significant performance improvements over the NJUDS, STERE, DES and NLPR datasets. The advantage of our method over the SSD dataset is not obvious due to its controversial human-labeled annotations. Some saliency ground truth annotations in the SSD dataset may be somewhat controversial: i.e., there are 80 images in total in the SSD dataset, and almost 21% (17/80) of them clearly contain multiple salient objects, while the saliency ground truth annotations of these images tend to regard only one of them as the salient one. To better understand this issue, we demonstrate several most representative cases in Fig. 7, where most of these images contain multiple salient objects, while only one of them is annotated as the salient object. In these cases, most of the SOTA approaches tend to focus on a single salient object; however, our method tends to detect more objects, and this is exactly why our method is capable of outperforming other approaches in terms of detection completeness over other four datasets. Though facing this problem, our method still belongs to the top-3 methods over the SSD dataset.

We have demonstrated the qualitative comparison results in Fig. 8. We may easily notice that the salient object detection results of the compared SOTA methods may frequently become degenerated when either the RGB component or the depth part partially fails to separate the salient object from its nonsalient nearby surroundings. For example, in the bottom row of Fig. 8, the salient object in the RGB component has exhibited a large uniqueness degree, yet it is difficult for the SOTA methods to achieve the correct saliency estimation over the depth channel, leading to an incomplete detection result.

Furthermore, as shown in the second row of Fig. 8, the depth component is extremely useful in this case; however, almost all of the compared methods are incapable of producing correct detection results due to the low contrast RGB component, which leads to an inappropriate complementary status between RGB and depth information.

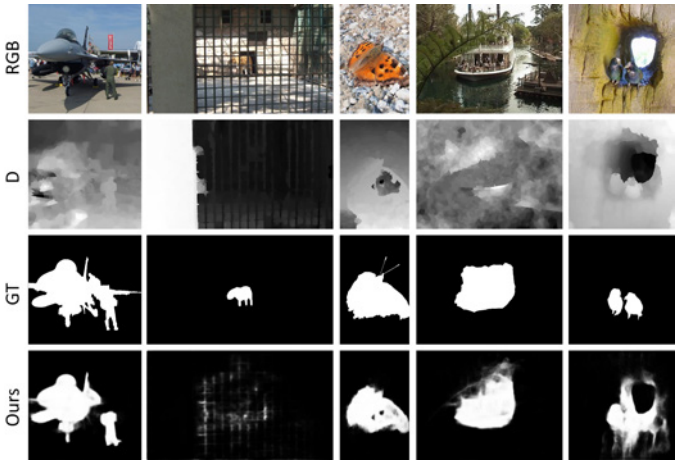

Fig. 9: Demonstration of some failure cases.

L. Implementation Details

We leverage the Caffe toolbox to implement our method. We train and test our model on a desktop computer with an NVIDIA GTX 1080 GPU (8 GB memory), an Intel Core I7-6700 CPU (4 cores with 8 threads, 3.40 GHz) and 32 GB RAM.

For a fair comparison, we follow an identical training and testing protocol adopted in [46]; i.e., our training set consists of 1400 stereo images from the NJUDS [47] dataset and 650 stereo images from the NLPR [39] dataset, and the rest of the data is used as the testing dataset.

Our training includes three steps: 1) we firstly train our 3 independent backbone nets by solely using the RGB component; 2) we fine-tune each subbranch using the newly formulated input data, i.e., DGB, RDB, and RGD; 3) we jointly fine-tune the triple-stream fusion network. For each step in the training phase, 5000 iterations with SGD backpropagation are required. The learning rate, weight decay, momentum and iter size are assigned as $\{$1e-7, 0.0005, 0.9, 10$\}$, respectively.

In the testing phase, our model requires almost $0.064s$ (15.6 FPS) to conduct salient object detection for a single 224×224 RGB-D image.

M. Limitations and Future Works

Although the proposed model can perform well on most cases, it might still encounter failure cases, where Fig. 9 has demonstrated several most representative ones. The main reason is that the depth information usually varies from scene to scene and its quality is mainly determined by both scene layout and sensing equipment. In facing a RGB-D image with low-quality depth, the corresponding depth saliency degenerates, making the fused RGB-D saliency even worse (see the middle three columns in Fig. 9). We name those depth channels, which

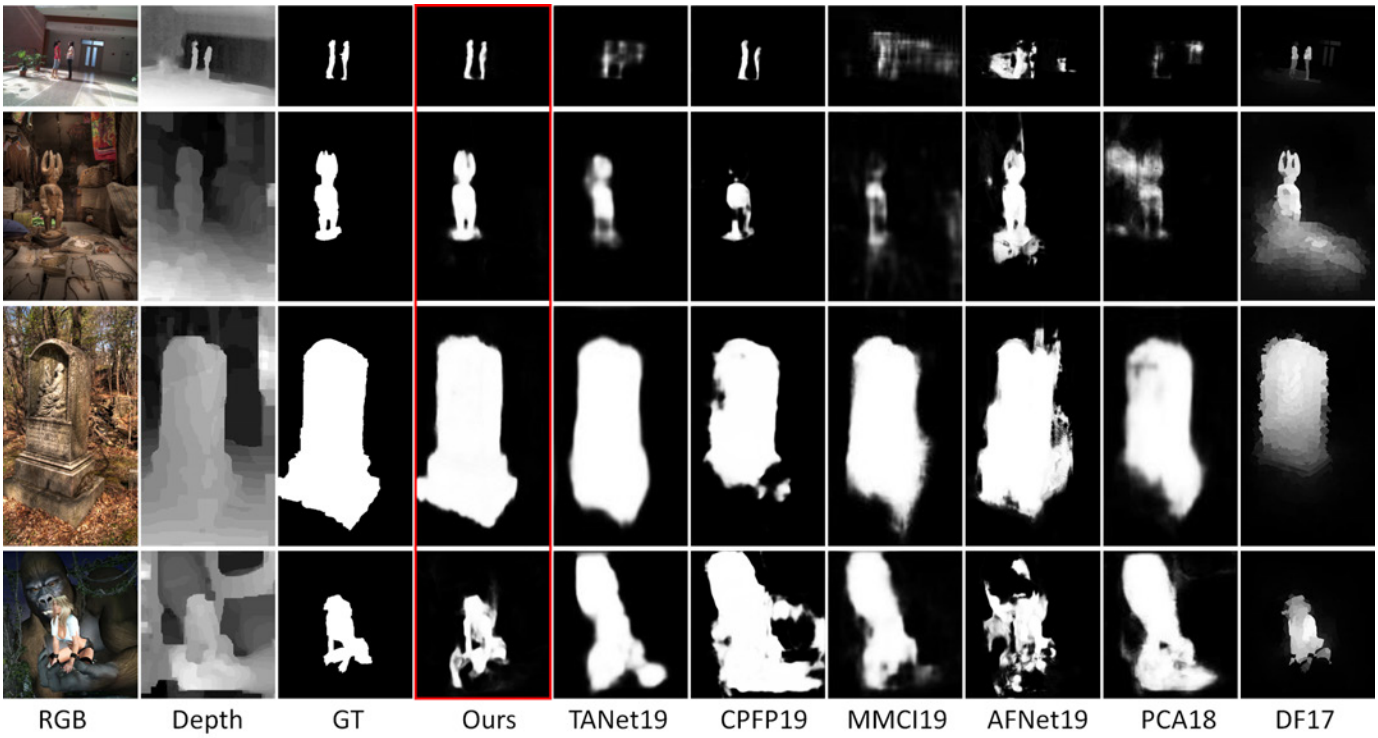

Fig. 8: Qualitative comparisons between our method and the other 6 most representative SOTA methods, including TANet [59], CPFP [46], MMCI [58], AFNet [38], PCA [57] and DF [33].

cannot help in separating salient objects from their non-salient nearby surroundings, as the "low-quality" ones.

In the future work, we are particularly interested in exploring novel schemes which might be able to convert the RGB-D fusion being depth-quality-aware. Thus, the fused RGB-D saliency maps might be capable of biasing towards the RGB information when the quality of depth channel cannot help in separating salient objects from their non-salient nearby surroundings, avoiding the side-effects induced by low-quality depth.

V. Conclusion

In this paper, we have developed a novel channel-wise fusion network to conduct multi-net and multi-level selective fusion for high-performance RGB-D salient object detection. To achieve this objective, we have designed a novel backbone network which receives our newly formulated input data to pursue an optimal complementary status between RGB and depth channels. Specifically, our backbone network is implemented based on the VGG16 network, which can be replaced by other high-performance networks, further improving the overall performance of our method. Then, we have proposed a novel triple-stream fusion network to ensure an optimal fusion state for each of our subnetworks in multi-level fashion. Moreover, we have conducted extensive quantitative evaluations to verify the effectiveness of our method.

Acknowledgments. This research is supported in part by the National Natural Science Foundation of China (No. 61802215, 61806106, 61672077 and 61532002), the Natural Science Foundation of Shandong Province (Nos. ZR2019BF011 and ZR2019QF009) and the National Science Foundation of the USA (Nos. IIS-1715985 and IIS-1812606).

References

[1] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3203–3212.

[2] Y. Fang, C. Zhang, H. Huang, and J. Lei, "Visual attention prediction for stereoscopic video by multi-module fully convolutional network," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5253–5265, 2019.

[3] K. Fu, D. Fan, G. Ji, and Q. Zhao, "Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[4] F. Sun, T. Kong, W. Huang, C. Tan, B. Fang, and H. Liu, "Feature pyramid reconfiguration with consistent loss for object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5041–5051, 2019.

[5] G. Ma, S. Li, C. Chen, A. Hao, and H. Qin, "Stage-wise salient object detection in 360 omnidirectional image via object-level semantical saliency ranking," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, pp. 1–1, 2020.

[6] S. Battiato, G. M. Farinella, G. Puglisi, and D. Ravi, "Saliency-based selection of gradient vector flow paths for content aware image resizing," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2081–2095, 2014.

[7] G. Hou, X. Zhao, Z. Pan, H. Yang, L. Tan, and J. Li, "Benchmarking underwater image enhancement and restoration, and beyond," *IEEE Access*, vol. 8, pp. 122 078–122 091, 2020.

[8] L. Zhang, L. Wang, and W. Lin, "Conjunctive patches subspace learning with side information for collaborative image retrieval," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3707–3720, 2012.

[9] G. Liu and D. Fan, "A model of visual attention for natural image retrieval," in *International Conference on Information Science & Cloud Computing Companion*, 2013, pp. 728–733.

[10] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[11] J. Guo, T. Ren, L. Huang, X. Liu, M. Cheng, and G. Wu, "Video salient object detection via cross-frame cellular automata," in *IEEE International Conference on Multimedia and Expo*, 2017, pp. 325–330.

[12] Y. Fang, Z. Chen, W. Lin, and C. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3888–3901, 2012.

[13] B. Yan, W. Tan, K. Li, and Q. Tian, "Codebook guided feature-preserving for recognition-oriented image retargeting," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2454–2465, 2017.

[14] C. Chen, S. Li, H. Qin, and A. Hao, "Real-time and robust object tracking in video via low-rank coherency analysis in feature space," *Pattern Recognition*, vol. 48, pp. 2885–2905, 2015.

[15] C. Chen, S. Li, A. Hao, and H. Qin, "Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis," *Pattern Recognition*, vol. 52, pp. 410–432, 2016.

[16] C. Chen, G. Wang, C. Peng, X. Zhang, and H. Qin, "Improved robust video saliency detection based on long-term spatial-temporal information," *IEEE Transactions on Image Processing*, vol. 29, pp. 1090–1100, 2019.

[17] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4819–4831, 2019.

[18] Y. Li, S. Li, C. Chen, A. Hao, and H. Qin, "Accurate and robust video saliency detection via self-paced diffusion," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1153–1167, 2019.

[19] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3156–3170, 2017.

[20] Y. Li, S. Li, C. Chen, A. Hao, and H. Qin, "A plug-and-play scheme to adapt image saliency deep model for video data," *IEEE Transactions on Circuits and Systems for Video Technology (TVCG)*, pp. 1–1, 2020.

[21] D. Fan, W. Wang, M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8554–8564.

[22] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi-supervised video object segmentation with super-trajectories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 985–998, 2019.

[23] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *IEEE International Conference on Computer Vision*, 2019, pp. 7223–7233.

[24] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, 2017.

[25] T. Zhou, D. Fan, M. Cheng, J. Shen, and L. Shao, "Salient object detection: A survey," *arXiv preprint arXiv:2008.00230*, 2020.

[26] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global contextaware progressive aggregation network for salient object detection," *arXiv preprint arXiv:2003.00651*, 2020.

[27] Y. Kong, J. Zhang, H. Lu, and X. Liu, "Exemplar-aided salient object detection via joint latent space embedding," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5167–5177, 2018.

[28] G. Ma, C. Chen, S. Li, C. Peng, A. Hao, and H. Qin, "Salient object detection via multiple instance joint re-learning," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 324–336, 2019.

[29] L. Zhang, J. Wu, T. Wang, A. Borji, G. Wei, and H. Lu, "A multistage refinement network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 3534–3545, 2020.

[30] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2303–2316, 2015.

[31] D. Fan, Z. Lin, Z. Zhang, M. Zhu, and M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, pp. 1–1, 2020.

[32] C. Li, R. Cong, S. Kwong, J. Hou, and Q. Huang, "Asif-net: Attention steered interweave fusion network for rgb-d salient object detection," *IEEE Transactions on Cybernetics (TCYB)*, pp. 1–1, 2020.

[33] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "Rgbd salient object detection via deep fusion," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2274–2285, 2017.

[34] Y. Piao, X. Li, M. Zhang, J. Yu, and H. Lu, "Saliency detection via depth-induced cellular automata on light field," *IEEE Transactions on Image Processing*, vol. 29, pp. 1879–1889, 2019.

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

[36] X. Zhou, H. Wen, R. Shi, H. Yin, and C. Yan, "Depth-guided saliency detection via boundary information," *Image and Vision Computing*, vol. early access, pp. 1–1, 2020.

[37] X. Zhou, G. Li, C. Gong, Z. Liu, and J. Zhang, "Attention-guided rgbd saliency detection using appearance information," *Image and Vision Computing*, vol. 95, pp. 1–10, 2020.

[38] N. Wang and X. Gong, "Adaptive fusion for rgb-d salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.

[39] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *European Conference on Computer Vision*, 2014, pp. 92–109.

[40] J. Ren, X. Gong, Y. Lu, W. Zhou, and M. Yang, "Exploiting global priors for rgb-d saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 25–32.

[41] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2343–2350.

[42] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 819–823, 2016.

[43] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, "An iterative co-saliency framework for rgbd images," *IEEE Transactions on Cybernetics*, vol. 49, no. 1, pp. 233–246, 2017.

[44] R. Shigematsu, D. Feng, S. You, and N. Barnes, "Learning rgb-d salient object detection using background enclosure, depth contrast, and top-down features," in *IEEE International Conference on Computer Vision*, 2017, pp. 2749–2757.

[45] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "Pdnet: Prior-model guided depth-enhanced network for salient object detection," in *IEEE International Conference on Multimedia and Expo*, 2019, pp. 199–204.

[46] J. Zhao, Y. Cao, D. Fan, M. Cheng, X. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[47] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Processing: Image Communication*, vol. 38, pp. 115–126, 2015.

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

[49] C. Zhu and G. Li, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *IEEE International Conference on Computer Vision*, 2017, pp. 3008–3014.

[50] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *International Conference on Internet Multimedia Computing and Service*, 2014, pp. 23–27.

[51] D. Fan, M. Cheng, J. Liu, S. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *IEEE International Conference on Computer Vision*, 2018.

[52] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *IEEE International Conference on Computer Vision*, 2017, pp. 4548–4557.

[53] D. Fan, C. Gong, Y. Cao, B. Ren, M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *International Joint Conference on Artificial Intelligence*, 2018, pp. 698–704.

[54] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3089–3098.

[55] J. Liu, Q. Hou, M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[56] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[57] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

[58] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection," *Pattern Recognition*, vol. 86, pp. 376–385, 2019.

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

[60] 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 Transactions on Cybernetics*, 2019.
[61] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *IEEE International Conference on Image Processing*, 2015, pp. 1115–1119.
[62] J. Guo, T. Ren, and J. Bei, "Salient object detection for rgb-d image via saliency evolution," in *IEEE International Conference on Multimedia and Expo*, 2016, pp. 1–6.
[63] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *IEEE International Conference on Computer Vision*, 2017, pp. 1509–1515.
[64] 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 Transactions on Image Processing*, vol. 26, no. 9, pp. 4204–4216, 2017.
[65] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, and L. Qing, "Stereoscopic saliency model using contrast and depth-guided-background prior," *Neurocomputing*, vol. 275, pp. 2227–2238, 2018.
[66] 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 Transactions on Cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2018.