Knowing Depth Quality In Advance: A Depth Quality Assessment Method For RGB-D Salient Object Detection

Xuehao Wang[1]   Shuai Li[1]   Chenglizhao Chen[2*]   Aimin Hao[1]   Hong Qin[3]

[1]State Key Laboratory of VRTS, Beihang University
[2]College of Computer Science and Technology, Qingdao University
[3]Stonybrook University

Abstract

Previous RGB-D salient object detection (SOD) methods have widely adopted deep learning tools to automatically strike a trade-off between RGB and D (depth), whose key rationale is to take full advantage of their complementary nature, aiming for a much-improved SOD performance than that of using either of them solely. However, such fully automatic fusions may not always be helpful for the SOD task because the D quality itself usually varies from scene to scene. It may easily lead to a suboptimal fusion result if the D quality is not considered beforehand. Moreover, as an objective factor, the D quality has long been overlooked by previous work. As a result, it is becoming a clear performance bottleneck. Thus, we propose a simple yet effective scheme to measure D quality in advance, the key idea of which is to devise a series of features in accordance with the common attributes of high-quality D regions. To be more concrete, we conduct D quality assessments for each image region, following a multi-scale methodology that includes low-level edge consistency, mid-level regional uncertainty and high-level model variance. All these components will be computed independently and then be assembled with RGB and D features, applied as implicit indicators, to guide the selective fusion. Compared with the state-of-the-art fusion schemes, our method can achieve a more reasonable fusion status between RGB and D. Specifically, the proposed D quality measurement method achieves steady performance improvements for almost 2.0% in general.

1. Introduction and Motivation

Salient object detection (SOD) aims to fast locate the most eye-attractive objects in a given scene [1, 2], and its downstream applications usually include object detection [3, 4], object segmentation [5, 6], image reconstruction [7, 8], visual tracking [9, 10] and video saliency detection [11–13]. Different to the existing SOD deep models using RGB information solely [14–16], the RGB-D SOD [17, 18], as the main foci of this paper, takes both RGB and D (depth) as input to make a complementary fusion for the SOD task.

In general, RGB-D SOD deep models follow a hypothesis that salient object should be located at a different D layer to the non-salient surroundings nearby [19]. Thus, most of the state-of-the-art (SOTA) approaches [20] have followed the bi-stream network architecture, the key methodology of which is to compute saliency clues over RGB channels and D channel respectively first and fuse these clues to obtain an RGB-D SOD result later. Since both RGB and D saliency clues can be easily computed via off-the-shelf deep models that follow the multi-level/multi-scale contrast computations, the fusion procedure is a critical factor for the overall RGB-D SOD performance.

In most cases, the widely-used selective fusion (Fig. 1-A) is capable of biasing its fused RGB-D saliency map towards either RGB channels or D channel to a certain extent. By taking numerous RGB-D saliency combinations as training input, the selective fusion models learn how to integrate its two individual inputs, which are respectively the output of RGB branch and D branch, via weighted pixel-wise operations. Though it can boost the overall performance indeed, the widely-used selective fusion scheme has one critical limitation in common: its unawareness of D quality easily results in various learning ambiguities, leading to a performance bottleneck eventually. For example, in the face of training instances (i.e., RGB-D images) with high-quality D, the

2

A. The Vanilla Bi-stream RGB-D SOD network

RGB-D Images → RGB Branch / D Branch → Fused RGB-D Saliency Map → Loss → GT

RGB | D | Sal | GT

The main learning task for the conventional RGB-D fusion networks is to learn how to selectively balance RGB and D. However, these methods easily encounter various learning ambiguities due to their less consideration of the D quality, leading to failure cases occasionally (e.g., Sal).

B. The Proposed Novel Method

RGB-D Images → RGB Branch / D Branch → Fused RGB-D Saliency Map → Loss → GT

D Quality Measurement → D Quality

RGB | D | Sal | GT

In sharp contrast, our method measures D quality in advance, which shrinks problem domain effectively and leads to a much simple learning task, producing much improved detection results.
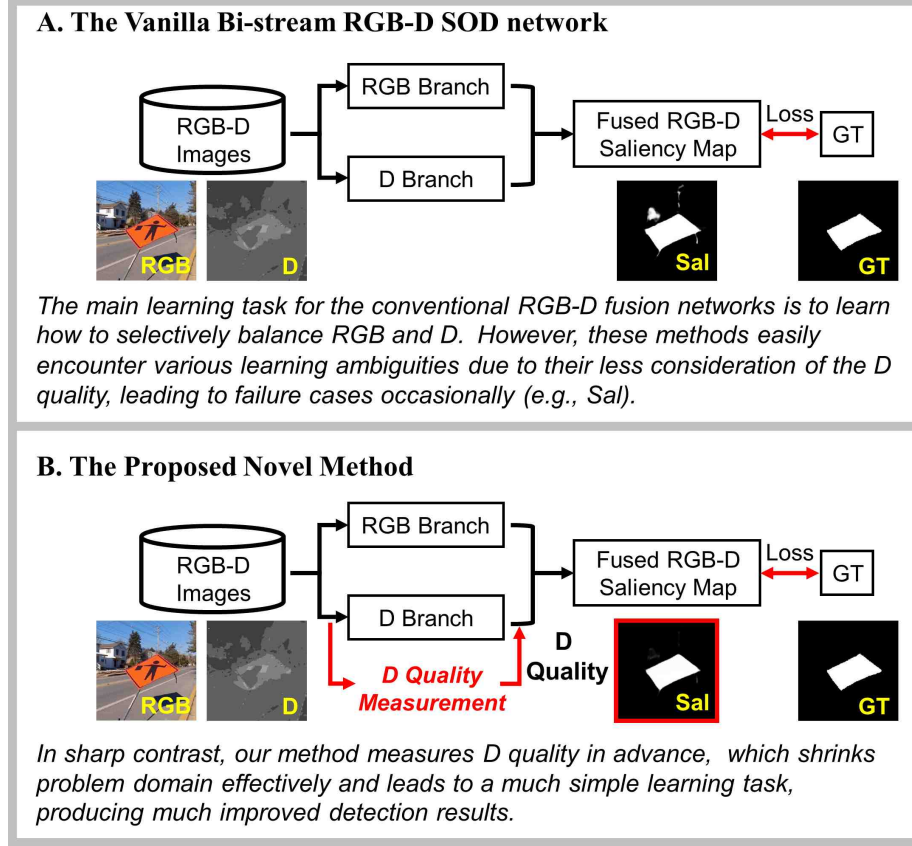
Figure 1: The main difference between the conventional methods and the proposed novel method.

learning scope of the selective fusion scheme should be focused on achieving an optimal combination of RGB and D, which is quite simple and easy in general; however, this learning task will become extremely complex and difficult when the training set contains RGB-D images with various D qualities, which usually correlates to a large problem domain with various learning ambiguities.

To conquer this limitation, we propose a simple yet effective scheme to measure D quality in advance (Fig. 1-B), the key idea of which is to devise a series of features in accordance with the common attributes of high-quality D image regions. To be more concrete, image regions with high-quality D usually have the following attributes:

1) Image regions with high-quality D should be capable of separating salient objects

3

from their non-salient surroundings nearby; based on this fact, we resort the edge consistency between RGB and D to measure D quality from a "low-level" perspective, which will be further detailed in Sec. 4.1.

2) In addition, the object-wise homogeneity in D values can constraint the objects' inner regions to be assigned with similar saliency values, which is critical for a complete SOD in face of a salient object that exhibits significant differences in its partial appearances; thus, we propose the regional-wise uncertainty to measure D quality in a "mid-level" way, which will be further explained in Sec. 4.2.

3) Most importantly, the D quality can be measured from the deep model itself implicitly, i.e., the fused RGB-D saliency can only get improved by using high-quality D, while low-quality D may degenerate the fused saliency; therefore, we shall conduct a "high-level" measurement, i.e., computing D quality via the performance variance between deep models that are respectively fed by the 3-dimensional RGB and the 4-dimensional RGB-D, which will be introduced in Sec. 4.3.

All these D quality features will be computed independently, and then be assembled to guide selective fusion between RGB and D. The salient contributions of this paper can be summarized as:

- As the first attempt, we have provided a deep insight into the D quality, which is a critical factor for the RGB-D SOD fusion performance, while it has long been overlooked by previous work.

- We have proposed a "multi-level" D quality measurement to adaptively guide RGB-D saliency fusion, which can effectively alleviate the learning ambiguities and achieve a much-improved SOD performance.

- We have conducted extensive quantitative evaluations to prove the effectiveness of the proposed method; we have conducted massive quantitative comparisons to show its performance superiority.
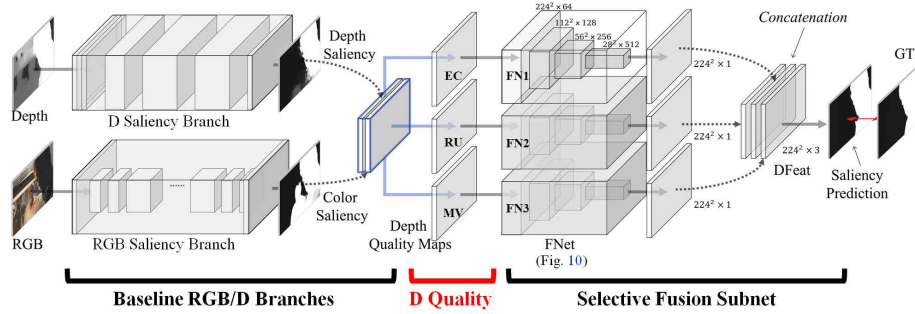
4

Figure 2: The overall network architecture of our method, where the "D Quality" is the main contribution of this paper.

- Both source code and results are publicly available at `https://github.com/XueHaoWang-Beijing/DQSF`, which may potentially be able to benefit the RGB-D SOD community in the future.

2. Related Work

The SOTA RGB-D SOD methods usually treat D as an additional image channel, the key rationale of which is to combine RGB saliency and D saliency simply, aiming for the improved overall performance. Following the vanilla bi-stream fusion methodology, Desingh et al. [21] compute the low-level saliency clues over RGB and D channels respectively, then combine these clues to obtain the RGB-D saliency. Similarly, Ren et al. [22] utilize the multiplicative based fusion to integrate three whole-map saliency features, including D saliency clues, RGB saliency clues and global appearance priors. Inspired by the phenomenon that salient objects are more likely to be located in front of the image backgrounds, Feng et al. [23] propose the depth orientation to measure D saliency. However, this method easily produces failure detections when the salient objects are not in front of the non-salient backgrounds.

With the rapid development of deep learning tools, the deep fusion-based SOTA methods are capable of biasing their fusion towards either RGB or D to a certain extent. Qu et al. [24] leverage the convolutional neural networks (CNNs) to selectively

fuse multiple low-level handcrafted saliency clues. Shigematsu et al. [25] extract multiple mid-level handcrafted features from depth channel to make the saliency fusion more robust. Zhu et al. [26] propose a depth-enhanced network, which consists of two subnetworks; i.e., one master network aims for RGB saliency computation, and the other makes full use of D saliency by integrating its deep features into the master network. Liu et al. [27] feed the concatenation of original depth channel and RGB channels into a single-stream recurrent convolution neural network based on the multi-scale and multi-level fusion. Chen et al. [28] conduct RGB-D saliency fusion via a newly designed multi-modal fusion network, which is capable of using multi-scale, multi-path and cross-modal interactions to promote RGB-D SOD performance. Cong et al. [29] measure the channel-wise importance in advance, and then use it to determine whether the RGB channels or the D channel should be biased during the fusion process. Liu et al. [30] add the color-stream features into the decoder network of depth stream to overcome the shortcomings of poor-quality depth images and then fuse the multi-modal results under the control of an adaptive gated fusion module. Zhao et al. [31] have mentioned the importance of D quality and introduced a novel RGB based contrast loss into the D stream, aiming to enhance the quality of D features.

3. Method Overview

We show the method overview in Fig. 2, which mainly consists of three components: 1) RGB/D baseline branches; 2) D quality; 3) Selective fusion subnet. Following the vanilla bi-stream structure, the first component includes two individual subbranches, i.e., one for the RGB saliency computation and the other for the D saliency computation, which can be any off-the-shelf deep model. Next, we resort three individual D quality feature maps, which are measured off-line, to guide the selective fusion between RGB branch and D branch (a.k.a. RGB saliency and D saliency). This component is the main foci of our paper, each subpart of which will be respectively detailed
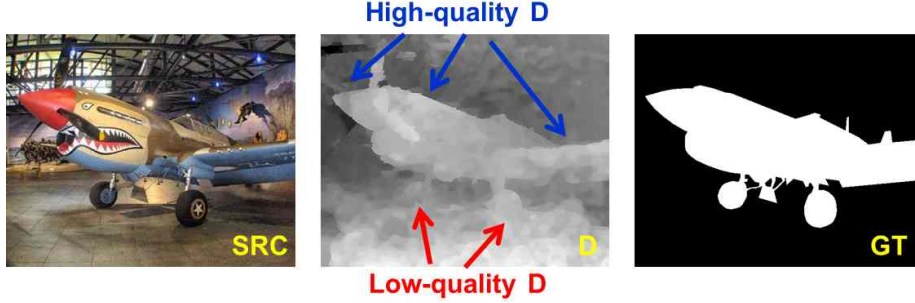
6

Figure 3: The demonstrations of image regions with different D qualities.

in Sec. 4. At last, we devise a selective fusion procedure, which is designed with three parallel UNet subbranches, to take full advantage of D quality feature maps, and avoid learning ambiguities.

4. Depth Quality Measurement

This section will investigate an effective scheme to conduct D quality assessments from a multi-scale perspective, which includes: 1) low-level edge consistency, 2) mid-level regional uncertainty and 3) high-level model variance.

4.1. Low-level Edge Consistency

Generally, there exists a significant common attribute of the image regions with high-quality D; i.e., it can easily separate salient objects from their non-salient surroundings nearby. For example, as shown in Fig. 3, some parts of the salient object can be easily separated from the non-salient regions by using D channel solely (e.g., the blue arrows), while some parts of the salient object may not be separated easily via D channel (e.g., the red arrows). In fact, such high-quality D regions are usually located around edges. In other words, those image regions near the edges that exhibit strong mutual consistency between RGB channels and D channel will have a large potential to be high-quality D regions. Inspired by this fact, we propose a simple yet effective scheme to locate image regions with high-quality D by measuring the low-level
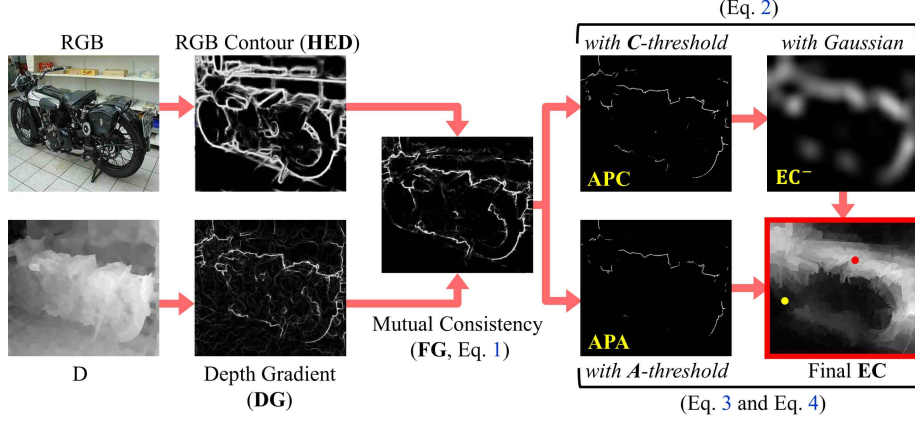
7

Figure 4: 1/3 pipeline of the D quality measurement: low-level edge consistency.

consistency between "RGB Contour" and "D Gradient" (**DG**), the method pipeline is demonstrated in Fig. 4.

Given a pair of RGB-D images ($\mathbf{I} \in \mathbb{R}^{W \times H}$, where W and H respectively represent the width and height), we use the off-the-shelf holistically-nested edge detection method (**HED** [32]) to obtain RGB contour maps. Compared with the conventional edge detection methods (e.g., Canny), the contour maps produced by **HED** can highlight object contours while suppressing those less relevant edges located in inner regions of the object.

The mutual consistency (**FG**) between RGB contour map (**HED**) and D gradient map (**DG**) can be simply formulated as Eq. 1, thus these pixels with high consistency degree will be interactively compressed.

$$\mathbf{FG} = \mathbf{DG} \odot \mathbf{HED}. \tag{1}$$

Here $\odot$ is the element-wise Hadamard product. We show the pictorial demonstration regarding the mutual consistency map (**FG**) in the middle column of Fig. 4.

Since the high-quality D regions tend to be located near those pixels with large **FG** values, we determine a subgroup of "anchor pixels" (**APC** in Fig. 4) by using a pre-

Figure 5: Qualitative demonstrations of the 1/3 D quality map (**EC**) using low-level edge consistency.

defined hard-threshold (T_c), and these anchor pixels will be used to coarsely locate the high-quality D regions via Eq. 2.

$$\mathbf{EC}^- = G\Big(pos(\mathbf{FG} - T_c)\Big),\tag{2}$$

where T_c is a pre-defined hard-threshold; G denotes the Gaussian smoothing (Gaussian Filtering) operation; pos is a function assigning its negative input into zero.

To produce a full regional-wise D quality map, we apply a novel spatial-weighting operation (Eq. 3) over $\mathbf{EC}^-$, which estimates D quality for the image regions that are not quite near edges. In fact, a typical spatial-weighting scheme should comprise the following two components, i.e., 1) feature similarity measurement (e.g., the exp component in Eq. 3, we implement it following the common thread mentioned in [13]); 2) spatial-weighting scope (i.e., the ϕ in Eq. 3 that is usually determined by a constant Euclidean distance). In sharp contrast, the spatial-weighting scope (ϕ) in our novel method is adaptively determined by a sub-group of most trustworthy anchor pixels via Eq. 4, in which these pixels are determined by using an aggressive hard-threshold (i.e., T_a, and $T_a \gg T_c$), see **APA** in Fig. 4. Specifically, we conduct the spatial-weighting

over super-pixels (SLIC [33]) to alleviate the computational burden.

$$\mathbf{EC}(sp_i) \leftarrow \frac{\sum\limits_{sp_j \in \phi} \mathbf{EC}^-(sp_j) \cdot exp\left(-\omega_1 \cdot \big|\big|c(sp_i), c(sp_j)\big|\big|_2\right)}{\sum\limits_{sp_j \in \phi} exp\left(-\omega_1 \cdot \big|\big|c(sp_i), c(sp_j)\big|\big|_2\right)}, \tag{3}$$

$$\phi : \Big|\Big|p(sp_i), p(sp_j)\Big|\Big|_2 \leq min\left\{\Big|\Big|sp_i, pos(\mathbf{FG} - \mathrm{T_a})\Big|\Big|_2\right\}, \tag{4}$$

where sp_i denotes the i-th superpixel; $\mathrm{T_a}$ is a predefined aggressive hard-threshold; ω_1 is a strength parameter; function $c(\cdot)$ and $p(\cdot)$ respectively return mean value and center position of their given RGB input. We demonstrate the final edge consistency map ($\mathbf{EC}$) in the bottom-right of Fig. 4 (marked with a red rectangle), in which we use a yellow/red dot to indicate some representative low-quality/high-quality D regions. More qualitative demonstrations of $\mathbf{EC}$ can be found in Fig. 5.

4.2. Mid-level Regional Uncertainty

Though the D quality map measured by the aforementioned edge consistency is generally trustworthy, it still has one major limitation. That is, the edge consistency based D quality measurement will become less trustworthy if the target image regions are not near object contours, e.g., the inner regions of salient objects. Thus, in this subsection, we will introduce a novel scheme (i.e., regional uncertainty) to complement the previous edge consistency, aiming for a more comprehensive D quality measurement.

Our regional uncertainty measurement is inspired by another common attribute of the high-quality D regions; i.e., the image regions belonging to an identical object or image area will potentially have high-quality D if their D values tend to be similar with each other. We show the method pipeline of our regional uncertainty measurement in Fig. 6, which mainly comprises two components: 1) localization prior; 2) regional-wise uncertainty.

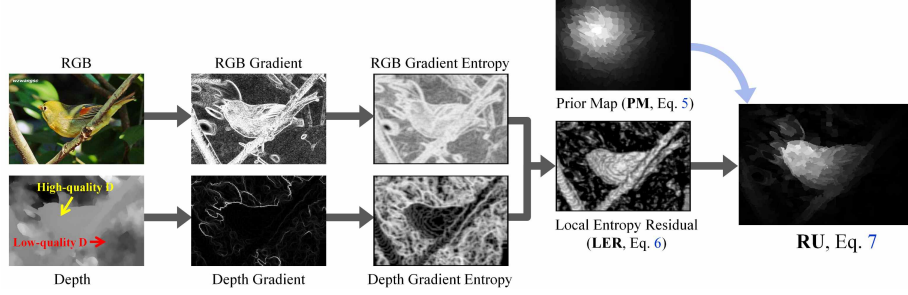Though the main interest of this subsection is to conduct D quality assessment for

Figure 6: 2/3 pipeline of the D quality measurement: mid-level regional uncertainty.

those regions that are not so near to object contours, we should precisely define the concept of "not so near" in advance, because it is less trustworthy and not necessary to conduct D quality assessment for the image regions that are far away from salient objects. Thus, we adopt the localization prior to indicate which regions will be the regions of our interest, i.e., the regions are "not so near" to object contours.

We formulate the localization prior ($\mathbf{PM} \in \mathbb{R}^{W \times H}$) as Eq. 5. For example, the localization prior of the i-th super-pixel is the summation of L2 spatial distances between it and each of the "most trustworthy anchor pixels" that have been introduced in the previous subsection, i.e., **APA**.

$$\mathbf{PM}(sp_i) = exp\left(- \omega_2 \cdot \sum_{l \leq PN} \left|\left| p(sp_i), \mathbf{HP}_l \right|\right|_2 \right), \tag{5}$$

where $\mathbf{PM}(sp_i)$ denotes the localization prior of the i-th superpixel; $\mathbf{HP} = \xi(\mathbf{FG} - \mathrm{T_a}) \in \mathbb{R}^{\mathrm{PN} \times 2}$, denoting the "most trustworthy anchor pixel" (i.e., **APA**), and PN is the total number of **APA**; **FG**, $\mathrm{T_a}$ and $p(\cdot)$ are identical to Eq. 4; function $\xi(\cdot)$ returns the coordinates of the non-negative elements; ω_2 is a weighting parameter. The qualitative demonstration toward **PM** can be found in the top-right of Fig. 6 (prior map).

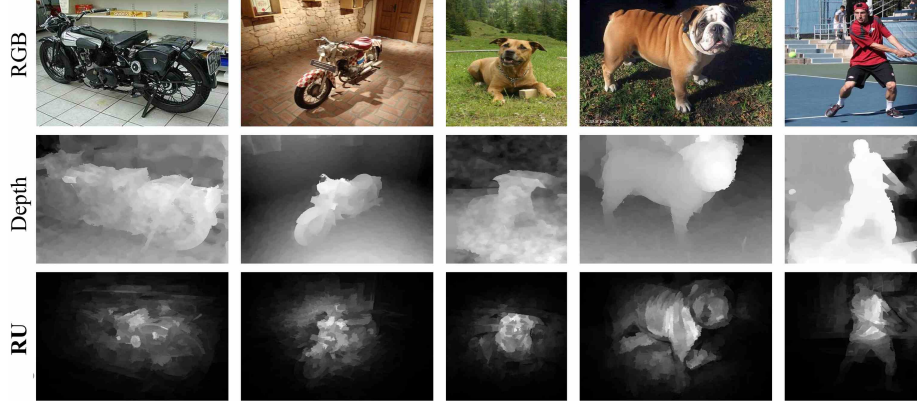We resort the non-local entropy residual (**LER**) to represent the regional uncer-

Figure 7: Qualitative demonstrations of the 2/3 D quality map (**RU**) using mid-level regional uncertainty.

tainty as Eq. 6.

$$\mathbf{LER}(p_i) = pos\Big(E(\mathbf{RGBG}, \phi_i) - E(\mathbf{DG}, \phi_i)\Big), \tag{6}$$

where **DG** and **RGBG** respectively represent D gradient map and RGB gradient map; function $E(\cdot, \cdot)$ returns the entropy value of the image region ϕ_i. The qualitative demonstration of **LER** can be viewed in Fig. 6, which usually exhibits large values in the regions with strong homogeneity in D channel yet with a large variance in RGB channels.

We utilize a simple multiplicative based fusion to integrate the previously computed localization prior and local entropy residual (**LER**) as our mid-level regional uncertainty based D quality map (**RU**), which can be formulated as Eq. 7).

$$\mathbf{RU} = \zeta\Big(\mathbf{LER} \odot \mathbf{PM}\Big), \tag{7}$$

where ζ denotes the common thread superpixel-wise spatial-weighting scheme that is identical to the spatial-weighting scheme mentioned in Eq. 3, the qualitative demonstration of which can be found in Fig. 7.
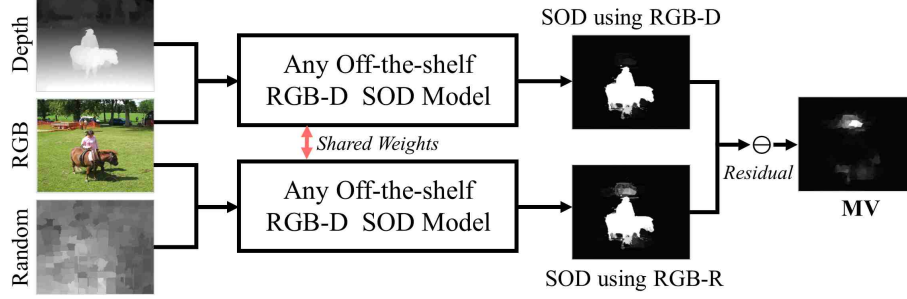
12

Figure 8: 3/3 pipeline of the D quality measurement: high-level model variance.

4.3. High-level Model Variance

In the previous subsections, we have introduced two explicit D quality features (i.e., **EC** and **RU**), following a handcrafted methodology. As another complementary component, in this subsection, we will introduce a novel implicit D quality measurement from the deep model perspective.

This component is inspired by the fact that RGB-D SOD models taking both RGB and D as input can significantly outperform the conventional SOD models using RGB information solely. Therefore, we propose the high-level model variance as another D quality measurement, the method pipeline can be represented by Fig. 8.

We use the variance/difference in SOD between two models to measure the implicit D quality, where these two SOD models share identical net architectures and weights yet are fed by different input channels, i.e., RGB-D and RGB-R (the last R represents the "*Random*" elements). We formulate the detailed model variance (**MV**) as Eq. 8.

$$\mathbf{MV} = \left| \Theta(\omega, \mathbf{RGBR}) - \Theta(\omega, \mathbf{RGBD}) \right|, \tag{8}$$

where $|\cdot|$ denotes the obsolete operation; function Θ represents the pre-trained RGB-D SOD model that receives 4-channel data as input, and ω denotes the learnable hidden parameters; **RGBD** denotes the 4-dimensional RGB-D image and **RGBR** denotes the newly formulated input data consisting of 3-channel RGB information and 1-channel
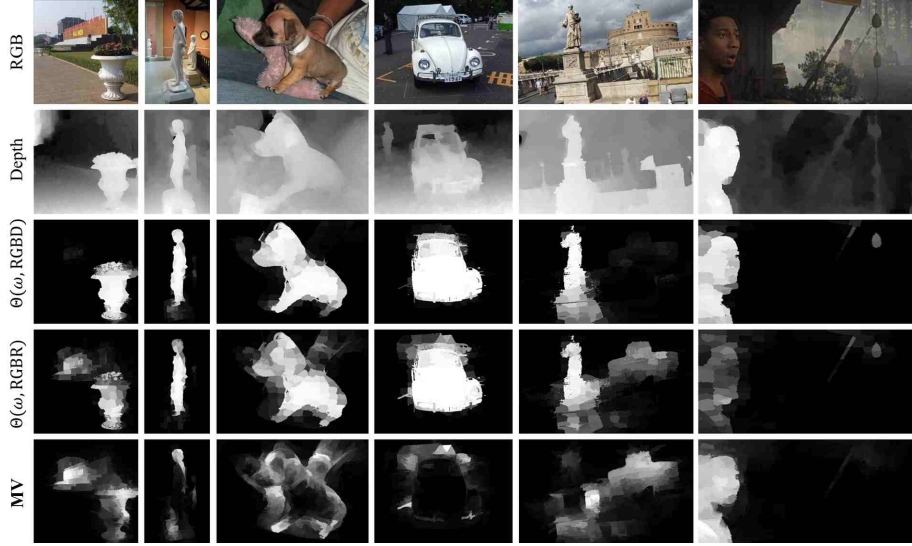
13

Figure 9: Qualitative demonstrations of the 3/3 D quality map (**MV**) using high-level model variance.

matrix with random noises. The **MV** highlights those D image regions that can benefit the RGB-D SOD task and, consequently, have a large probability to be the high-quality cases. The qualitative demonstrations of **MV** can be found in Fig. 9.

5. Saliency Fusion Guided by D Quality Features

5.1. Fusion Network Overview

Since the saliency computation over depth channel is theoretically similar to that over RGB channels, our main foci here is to investigate how to make full use of the previously obtained D quality features (i.e., **EC**, **RU** and **MV**) for the RGB-D saliency fusion.

As shown in Fig. 2, our network mainly consists of three components: 1) RGB/D baseline branches; 2) D quality; 3) Selective fusion subnet. The RGB/D branches can be any off-the-shelf deep models, in which we adopt the off-the-shelf PoolNet [34] as the RGB saliency subbranch, and adopt the pre-trained CPFP [31] as the D saliency subbranch, where the CPFP is pre-trained using the same RGB-D training set as our
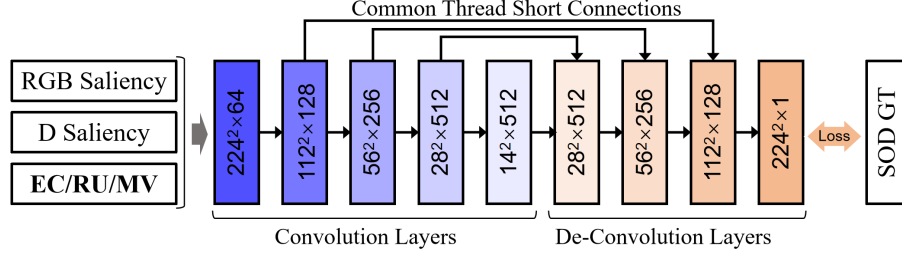
14

Figure 10: Network architecture of the proposed FNet; the relationship between the FNet and the whole RGB-D SOD net can be found in Fig. 2.

method.

The RGB saliency map and D saliency map will be respectively combined with each of the D quality features. Thus, the input of the fusion subnet that includes three parallel branches with an identical structure (i.e., FN1, FN2 and FN3) can be represented as $\{RGBSal, DSal, \mathbf{EC}\}$, $\{RGBSal, DSal, \mathbf{RU}\}$ and $\{RGBSal, DSal, \mathbf{MV}\}$. The fusion subnet (FNet) will be detailed in the next subsection. It is worth mentioning that a more complex fusion network will lead to better performance, though, we implement it using a lightweight designed structure mainly because this issue is beyond our main interest.

5.2. Fusion Subnet

As shown in Fig. 10, each subbranch of the FNet (i.e., FN1, FN2 and FN3, Fig. 2) follows the classic UNet structure to make full use of the multi-scale deep features of the precedent encoder layers, which iteratively integrates the deep features at different encoder layers into each decoder layer.

We use "DFeat" to represent the output of FNets (i.e., **FN1**, **FN2** and **FN3**), which can be detailed as Eq. 9.

$$
\begin{aligned}
\mathbf{DFeat} = FN1(\mathbf{CSal}, \mathbf{DSal}, \mathbf{EC}) \\
\cup FN2(\mathbf{CSal}, \mathbf{DSal}, \mathbf{RU}) \\
\cup FN3(\mathbf{CSal}, \mathbf{DSal}, \mathbf{MV}),
\end{aligned}
\tag{9}
$$

15

where the operator $\cup$ denotes the feature concatenate operation, and the obtained **DFeat** follow the formulation as **DFeat** $\in \mathbb{R}^{224 \times 224 \times \{1+1+1\}}$. The final saliency prediction can be obtained by using a convolutional operation (with 1×1 kernel) over the **DFeat**.

6. Experiments and Evaluations

6.1. Datasets

We have evaluated the proposed method on five public RGB-D benchmark datasets, which are listed below.

NJUDS [35]: This dataset contains 1,985 stereo image pairs, which are gathered from the internet, photographs and stereo movies with optical flow method; It consists of both simple and complex scenes. **NLPR** [20]: This dataset contains 1,000 images with the depth information captured by Microsoft Kinect in both indoor and outdoor scenes. It is more challenging because its scenes consist of multiple salient objects. **DES** [36]: This dataset is also called RGBD135 which contains 135 stereo images captured by Microsoft Kinect in seven indoor scenes. Most of the scenes have a single salient object. **LFSD** [37]: This dataset contains 100 stereo images with the depth information captured by Lytro light field camera. There is no clear boundary between the foreground and background regions in its depth channel. **STERE** [38]: This dataset is also named SSB. It contains 1,000 binocular images captured from both indoor and outdoor scenes.

6.2. Evaluation Metrics

We use F-measure [39], S-measure [40], E-measure [41] and MAE value to evaluate our performance. The F-measure is related to precision rate and recall rate. Given a predicted saliency map, we perform binary segmentation with a hard threshold T. If the obtained foreground is consistent with the ground truth mask, it is deemed as successful detection, and the final precision-recall curves are obtained by varying T from 0 to 255. As the recall rate is inversely proportional to the precision rate, the tendency

of the trade-off between precision and recall can truly indicate the overall detection performance.

F-measure is an important performance indicator when precision rate conflict with recall rate, and it can be computed as Eq. 10, which shows the balance between precision rate and recall rate.

$$\text{F-measure} = \frac{(\beta^2 + 1) \times \text{PRE} \times \text{REC}}{\beta^2 \times \text{PRE} + \text{REC}}, \tag{10}$$

where PRE represents the average precision rate, REC represents the average recall rate, and $\beta^2 = 0.3$ to balance the precision rate and the recall rate.

S-measure is also called Structure-measure [40]. The novel evaluation focuses on the region-wise and object-wise structural similarities, which is more similar to the human visual system. It can be formulated as:

$$\text{S-measure} = \alpha \times S_o + (1 - \alpha) \times S_r, \tag{11}$$

where we set $\alpha = 0.5$ to balance the region-aware (S_r) and object-aware (S_o) structural similarity.

E-measure is also named Enhanced-alignment Measure [41]. It combines the pixel-level evaluation (like F-measure) with image-level evaluation (like S-measure) to make a great improvement than other meta-measures. The formulation of this measure is shown as:

$$\text{E-measure} = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \theta_{\text{FM}}(x, y) \tag{12}$$

where w and h represent the width and height of the image respectively. θ is an enhanced alignment matrix [41] focused on the pixel-level matching and image-level statistics. FM represents the foreground map.

The MAE is defined as:

$$\text{MAE} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\text{SM}(x,y) - \text{GT}(x,y)|, \tag{13}$$

where W and H respectively represent the image width and image height; SM represents the estimated saliency map and GT denotes the ground truth.

6.3. Implementation Details

Our training set contains 2050 RGB-D images, including 1400 images from NJUDS dataset and 650 images from NLPR dataset. All these images are selected the same as [31] for a fair comparison. The testing dataset consists of the rest images. As for our **MV** feature (Eq. 8), we train the FNet (Fig. 10) for 10K iterations with 4-channel input (RGB-D) over the entire training set.

The parameters we mentioned in Sec. 4 will be detailed as follows: we assign the strength parameter ω_1 (Eq. 3), ω_2 (Eq. 5), the superpixel numbers in Eq. 3 and Eq. 5 respectively as $\{0.01, 7, 400, 1000\}$. Also, the conservative hard-threshold T_c (Eq. 2) and the aggressive hard-threshold T_a (Eq. 4) are respectively set as $\{20, 30\}$ times of the average of **FG** (Eq. 1). Particularly, we use the Gaussian smoothing (Gaussian Filtering) twice (Eq. 2) to initialize the regional-wise depth quality, in which the Gaussian parameters were respectively set to $\{80,25\}$ and $\{20,20\}$.

We optimize the entire network by using Stochastic Gradient Descent (SGD) with a moment 0.9, weight decay 0.005, iter size 10 and learning rate 1e-7.

6.4. Component evaluation

To validate the effectiveness of our method, we perform the component evaluation via S-measure, F-measure and MAE over the NJUDS testing dataset, see details in Table 1. As one of our baseline sub networks, the "Baseline-C" (PoolNet [34]) exhibits the worst performance in Table 1. Benefited by the usage of depth channel, the "Baseline-D" (CPFP [31]), which is another baseline subnetwork adopted in our
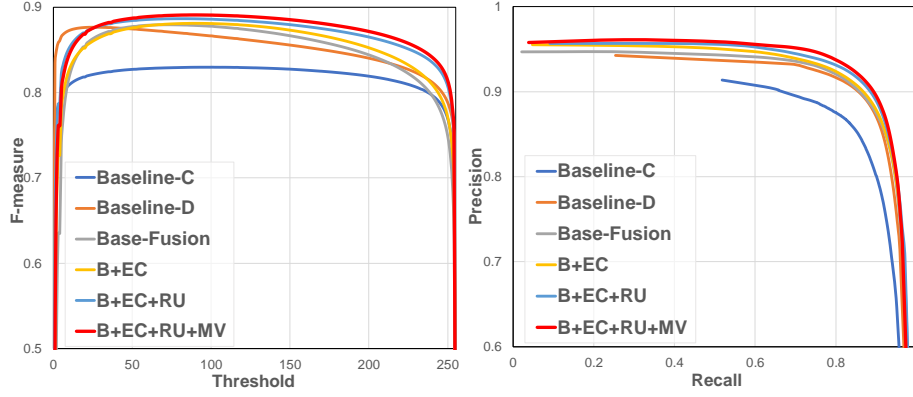
18

Figure 11: Precision-recall and F-measure curves of different combinations of key components.

Table 1: Quantitative evaluations regarding different combinations using various key components.

	Sm ↑	meanF ↑	maxF ↑	MAE ↓
Baseline-C	.843	.817	.877	.077
Baseline-D	.878	.850	.830	.053
Base-fusion(B)	.882	.842	.879	.061
B+EC	.884	.851	.882	.056
B+EC+RU	.887	.863	.885	.052
B+EC+RU+MV	**.892**	**.867**	**.891**	**.051**

method, exhibits a much-improved detection performance. The overall performance can be significantly improved by using our selective fusion subnetwork, see the "Base-fusion" in Table 1. Also, we may easily notice a significant performance improvement after integrating the edge-consistency-based D quality feature (**EC**, Sec. 4.1) into the base-fusion network, see the "B+EC" in Table 1. Then, the overall performance can be further improved by further using the regional-uncertainty-based D quality feature (**RU**, Sec. 4.2) and model-variance-based D quality feature (**MV**, Sec. 4.3), i.e, "B+EC+RU" and "B+EC+RU+MV", showing the effectiveness of the proposed D quality measurements. We show PR and F-measure curves of different combinations using various key components in Fig. 11. We can observe that the model with D quality features (marked as **B+EC+RU+MV**) achieves the best performance. The qualitative
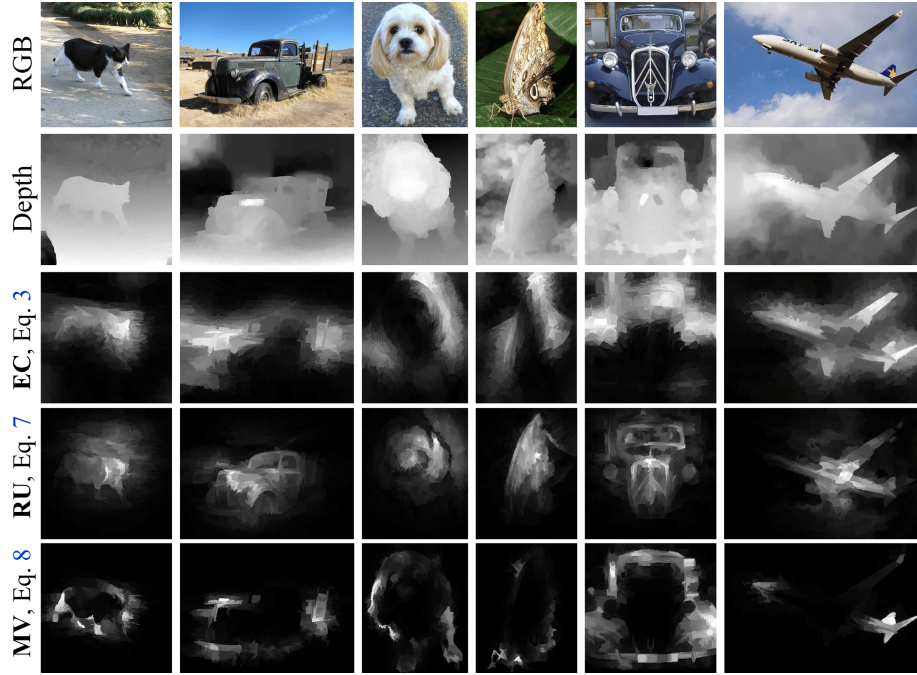
Figure 12: The overall demonstrations of all D quality features, including **EC**, **RU** and **MV**.

demonstrations of our D quality feature maps are shown in Fig. 12, and these feature maps are complementary with each other in general.

6.5. Performance comparisons to the SOTA methods

In this section, we compare our method with 13 other SOTA approaches, including CPFP [31], TANet [42], MMCI [28], AFNet [43], PCF [44], CTMF [45], CDB [46], DF [24], MDSF [47], CDCP [48], SE [49], DCMC [29], LBE [23]. For objective comparisons, all quantitative evaluations are conducted by using the source codes provided by the authors with parameters unchanged. The detailed quantitative results can be found in Table 2. Also, we provide the qualitative comparisons in Fig. 13, in which our method demonstrates three prominent advantages than these SOTA methods, i.e., 1) more complete detection, 2) rich in saliency details and 3) avoid negative effects induced by low-quality D. Moreover, for those images with high-quality D, our method can still outperform other SOTA methods.

20

Table 2: Comparison of quantitative results including F-measure (larger is better), E-measure (larger is better) , S-measure (larger is better) and MAE (smaller is better).

Set	Metric	LBE [23]	DCMC [29]	SE [49]	CDCP [48]	MDSF [47]	DF [24]	CDB [46]	CTMF [45]	PCF [44]	AFNet [43]	MMCI [28]	TANet [42]	CPFP [31]	Ours
NJUDS	**Sm** ↑	.695	.686	.664	.669	.748	.763	.624	.849	.877	.772	.858	.878	.878	**.892**
	adpE ↑	.791	.791	.772	.747	.812	.835	.745	.864	.896	.846	.878	.893	.895	**.910**
	maxE ↑	.803	.799	.813	.741	.838	.864	.742	.913	.924	.853	.915	.925	.923	**.928**
	adpF ↑	.740	.717	.734	.624	.757	.784	.648	.788	.844	.768	.812	.844	.837	**.856**
	meanF ↑	.606	.556	.583	.595	.628	.650	.482	.779	.840	.764	.793	.841	.850	**.867**
	maxF ↑	.748	.715	.748	.621	.775	.804	.648	.845	.872	.775	.852	.874	.877	**.891**
	MAE ↓	.153	.172	.169	.180	.157	.141	.203	.085	.059	.100	.079	.060	.053	**.051**
STERE	**Sm** ↑	.660	.731	.708	.713	.728	.757	.615	.848	.875	.825	.873	.871	.879	**.897**
	adpE ↑	.749	.831	.825	.796	.830	.838	.808	.864	.897	.886	.901	.906	.903	**.919**
	maxE ↑	.787	.819	.846	.786	.809	.847	.823	.912	.925	.887	.927	.923	.925	**.932**
	adpF ↑	.595	.742	.748	.666	.744	.742	.713	.771	.826	.807	.829	.835	.830	**.857**
	meanF ↑	.501	.590	.610	.638	.527	.617	.489	.758	.818	.806	.813	.828	.841	**.861**
	maxF ↑	.633	.740	.755	.664	.719	.757	.717	.831	.860	.823	.863	.861	.874	**.888**
	MAE ↓	.250	.148	.143	.149	.176	.141	.166	.075	.064	.068	.060	.051	.051	**.048**
DES	**Sm** ↑	.703	.707	.741	.709	.741	.752	.645	.863	.842	.770	.848	.858	.872	**.879**
	adpE ↑	.911	.849	.852	.816	.869	.877	.868	.911	.912	.874	.904	.919	.927	**.944**
	maxE ↑	.890	.773	.856	.811	.851	.870	.830	**.932**	.893	.881	.928	.910	.923	.931
	adpF ↑	.796	.702	.726	.625	.744	.753	.729	.778	.782	.730	.762	.795	.829	**.864**
	meanF ↑	.576	.542	.617	.585	.523	.604	.502	.756	.765	.713	.735	.790	.824	**.831**
	maxF ↑	.788	.666	.741	.631	.746	.766	.723	.844	.804	.729	.822	.827	.846	**.863**
	MAE ↓	.208	.111	.090	.115	.122	.093	.100	.055	.049	.068	.065	.046	.038	**.036**
NLPR	**Sm** ↑	.762	.724	.756	.727	.805	.802	.629	.860	.874	.799	.856	.886	.888	**.900**
	adpE ↑	.855	.786	.839	.800	.812	.868	.809	.869	.916	.884	.872	.916	.924	**.938**
	maxE ↑	.855	.793	.847	.820	.885	.880	.791	.929	.925	.879	.913	**.941**	.932	.938
	adpF ↑	.736	.614	.692	.608	.665	.744	.613	.724	.795	.747	.730	.796	.823	**.858**
	meanF ↑	.736	.543	.624	.609	.649	.664	.422	.740	.802	.755	.737	.819	.840	**.855**
	maxF ↑	.745	.648	.713	.645	.793	.778	.618	.825	.841	.771	.815	.863	.867	**.884**
	MAE ↓	.081	.117	.091	.112	.095	.085	.114	.056	.044	.058	.059	.041	.036	**.034**
LFSD	**Sm** ↑	.736	.753	.698	.717	.700	.791	.520	.796	.794	.738	.787	.801	.828	**.844**
	adpE ↑	.770	.842	.784	.780	.817	.844	.703	.851	.842	.810	.840	.845	.867	**.883**
	maxE ↑	.804	.856	.840	.786	.826	.865	.774	.865	.835	.815	.839	.847	.872	**.884**
	adpF ↑	.708	.816	.778	.697	.799	.806	.682	.782	.792	.742	.779	.794	.813	**.839**
	meanF ↑	.611	.655	.640	.680	.521	.679	.376	.756	.761	.735	.722	.771	.811	**.820**
	maxF ↑	.726	.817	.791	.703	.783	.817	.682	.791	.779	.744	.771	.796	.826	**.839**
	MAE ↓	.208	.155	.167	.167	.190	.138	.218	.119	.112	.133	.132	.111	.088	**.086**

As shown in Table 2, the F-measure of our method respectively achieves 1.6%, 1.6%, 1.8%, 2.0% and 1.6% improvements over the adopted datasets respectively. Meanwhile, our method consistently outperforms the SOTA methods in S-measure, E-measure and MAE as well. We also compare our model with other representative approaches in terms of PR and F-measure curves. As can be seen in Fig. 14 and Fig. 15, our model performs better than all the other approaches. Specifically, because the proposed edge consistency D quality measurement is developed on the gradient space, the depth-sensing equipment may directly affect the overall performance. Consequently, we can easily notice that our method performs the best in NLPR dataset (Microsoft Kinect, which can provide high-quality D) and the worst in LFSD dataset (Lytro light field camera, which can only provide low-quality D maps).
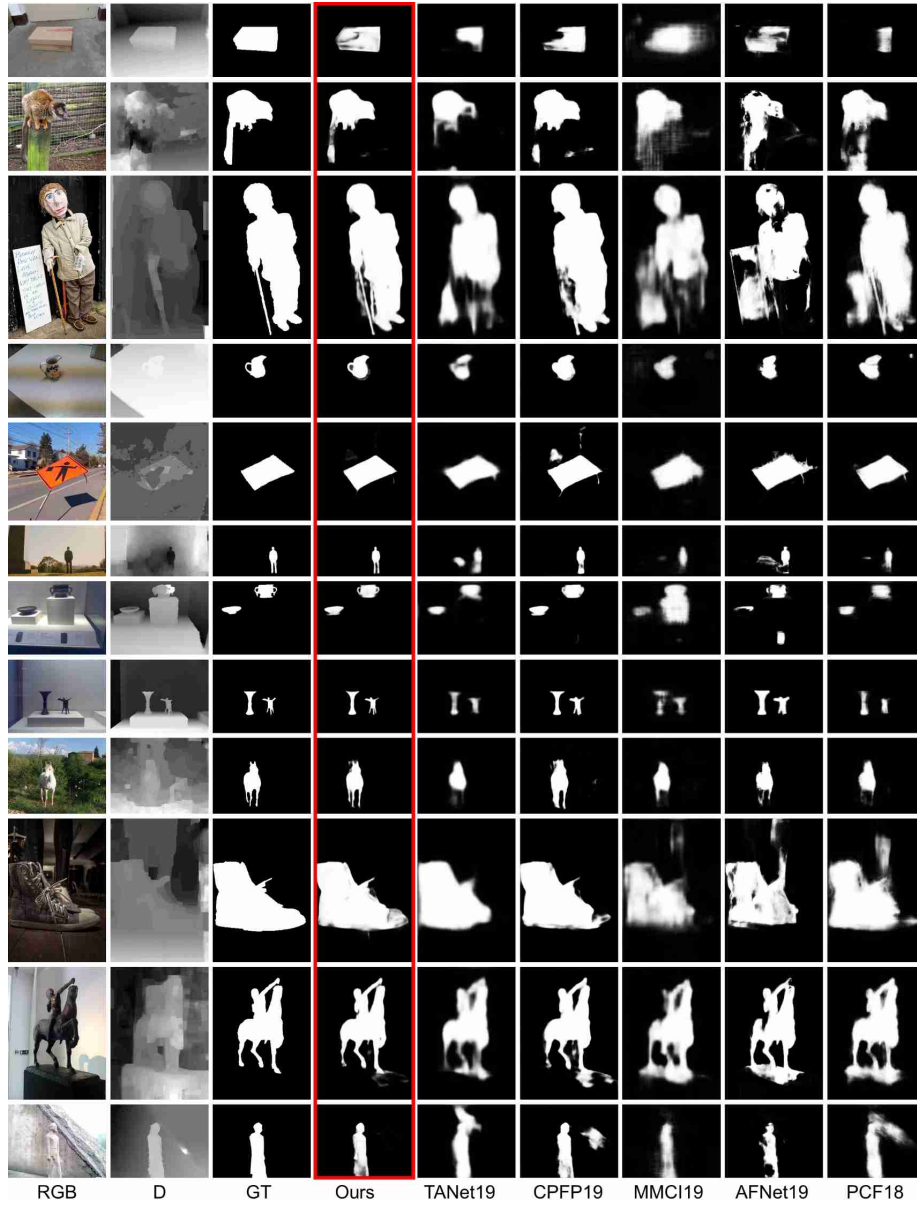
Figure 13: Qualitative comparisons to the SOTA methods. The qualitative comparisons listed here include TANet [42], CPFP [31], MMCI [28], AFNet [43] and PCF [44].
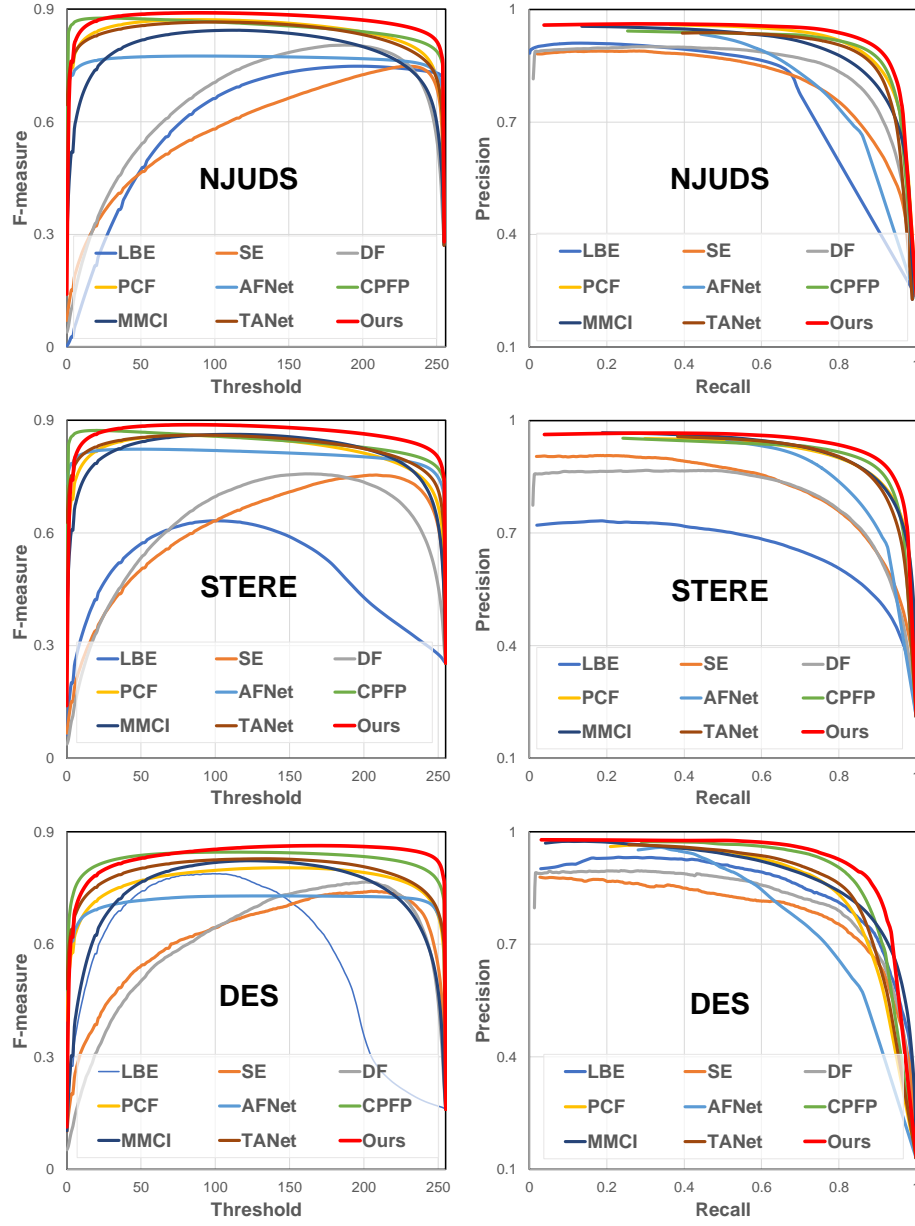
Figure 14: Quantitative comparisons on five popular datasets (3/5) in terms of the PR curves and F-measure curves.

7. Limitation

Our fusion network is lightweight designed, whose model size is quite smaller than other representative methods, taking more than 15% reduction. Its test time also
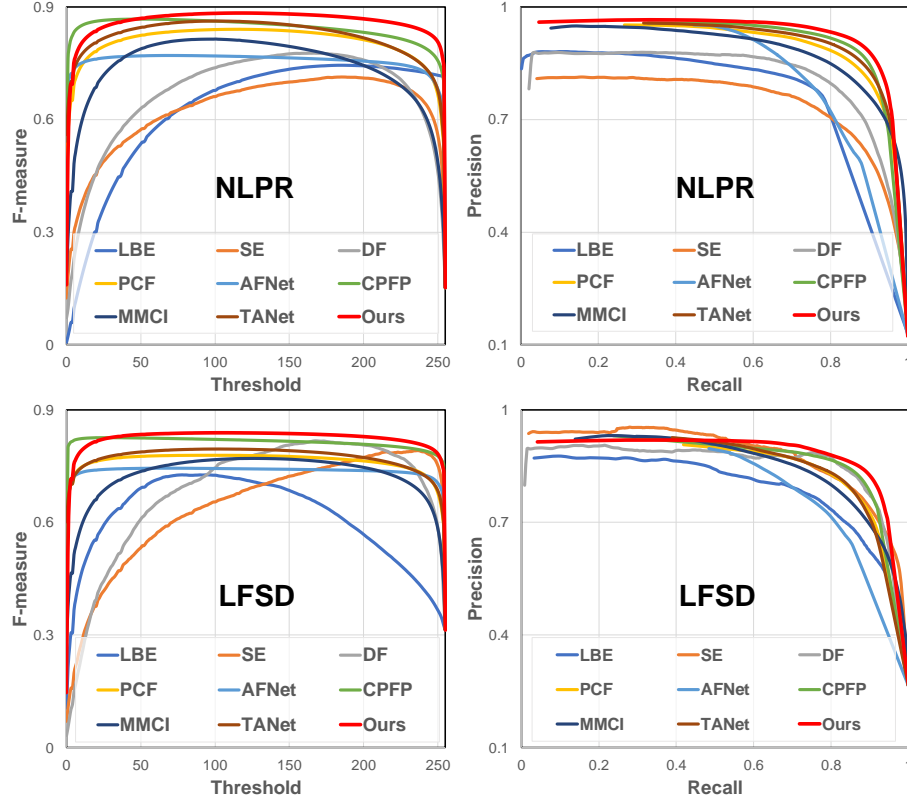
Figure 15: Continued quantitative comparisons on five popular datasets (2/5) in terms of the PR curves and F-measure curves.

achieves a leading performance. The quantitative comparisons can be seen in Table 3. However, our method is generally time-consuming due to the handcrafted D quality measurement features. On a desktop computer with i7-6700 3.40GHz CPU, GTX 1080 GPU and 32GB RAM, it takes almost 0.240s to compute our depth quality-aware features (4.2 FPS with CPU) and another 0.062s (16 FPS with GPU) to make the final saliency prediction. Thus, our method needs a total of 0.302s to perform SOD for a 224×224 RGB-D image. Also, our method may produce failure detection if both RGB and D are incapable of separating the salient object from its non-salient surroundings nearby.

Table 3: Comparison of model size and test time for each RGB-D image pair with other representative methods.

Model	PCF [44]	MMCI [28]	TANet [42]	CPFP [31]	Ours
Size (MB)	533.6	929.7	951.9	278.1	**235.0**
Time (ms)	65.5	**51.2**	70.4	170.0	62.1

8. Conclusions and Future Work

In this paper, we have proposed a novel D quality assessment solution to conduct "quality-aware" SOD for RGB-D images. The SOTA methods easily produce incorrect detections in the face of images with low-quality D. To conquer it, we have devised three novel features (i.e., **EC**, **RU** and **MV**) to measure the D quality before performing RGB-D saliency fusion. Meanwhile, we have devised an effective and lightweight designed fusion network to take full use of these D quality features during performing selective RGB-D fusion. The proposed idea regarding the D quality assessments will have a large potential to benefit the RGB-D SOD community in the future.

As for our near future work, we are particularly interested in developing a novel end-to-end depth quality assessment network, which is capable of measuring the depth quality very fast within a full-automatic manner. Moreover, we may use D quality maps to complete and enhance D channels.

References

References

[1] 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.

[2] 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.

[3] X. Wu, D. Sahoo, and S. C. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, 2020.

[4] K. Fu, T. Zhang, Y. Zhang, and X. Sun, "Oscd: A one-shot conditional object detection framework," *Neurocomputing*, p. early access, 2020.

[5] C. Hao, Y. Chen, Z. Yang, and E. Wu, "Higher-order potentials for video object segmentation in bilateral space," *Neurocomputing*, vol. 401, pp. 28–35, 2020.

[6] Z. Yang, Y. Ma, J. Lian, L. Zhu *et al.*, "Saliency motivated improved simplified pcnn model for object segmentation," *Neurocomputing*, vol. 275, pp. 2179–2190, 2018.

[7] D. Li, Q. Wang, and Y. Shen, "Multi-variable intelligent matching pursuit algorithm using prior knowledge for image reconstruction by l0 minimization," *Neurocomputing*, vol. 207, pp. 548–559, 2016.

[8] C. Peng, C. Chen, Z. Kang, J. Li, and Q. Cheng, "Res-pca: A scalable approach to recovering low-rank matrices," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7317–7325.

[9] Y. Zhang, X. Gao, Z. Chen, H. Zhong, and S. C. Hoi, "Learning weighted multi-model correlation filters for visual tracking," *Neurocomputing*, p. early access, 2020.

[10] Y. Du, Y. Yan, S. Chen, and Y. Hua, "Object-adaptive lstm network for real-time visual tracking with adversarial data augmentation," *Neurocomputing*, vol. 384, pp. 67–83, 2020.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] L. Huang, G. Li, Y. Li, and L. Lin, "Lightweight adversarial network for salient object detection," *Neurocomputing*, vol. 381, pp. 130–140, 2020.

[16] A. Dakhia, T. Wang, and H. Lu, "A hybrid-backward refinement model for salient object detection," *Neurocomputing*, vol. 358, pp. 72–80, 2019.

[17] 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.

[18] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *IEEE International Conference on Computer Vision*, 2019.

[19] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based rgb-d saliency detection via cross-view transfer and multi-view fusion," *IEEE Transactions on Cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2018.

[20] 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.

[21] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth." in *British Machine Vision Conference*, 2013.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] C. Zhu, X. Cai, K. Huang, T. Li, and G. Li, "Pdnet: Prior-model guided depth-enhanced network for salient object detection," *IEEE International Conference on Multimedia and Expo*, 2018.

[27] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient object detection for rgb-d image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, 2019.

[28] 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.

[29] 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, 2017.

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

[31] 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.

[32] S. Xie and Z. Tu, "Holistically-nested edge detection," in *IEEE International Conference on Computer Vision*, 2015, pp. 1395–1403.

[33] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk *et al.*, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

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

[35] 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.

[36] 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, p. 23.

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

[38] 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.

[39] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 248–255.

[40] 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.

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

[42] 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.

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

[44] 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.

[45] 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, 2017.

[46] 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.

[47] 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.

[48] 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.

[49] 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.