MVSalNet:Multi-View Augmentation for RGB-D Salient Object Detection

Jiayuan Zhou[1], Lijun Wang[1⋆], Huchuan Lu[1,2], Kaining Huang[3], Xinchu Shi[3], and Bocong Liu[3]

[1] Dalian University of Technology
[2] Peng Cheng Laboratory
[3] Meituan
zjy@mail.dlut.edu.cn
{ljwang,lhchuan}@dlut.edu.cn
{huangkaining,shixinchu,liubocong}@meituan.com

Abstract. RGB-D salient object detection (SOD) enjoys significant advantages in understanding 3D geometry of the scene. However, the geometry information conveyed by depth maps are mostly under-explored in existing RGB-D SOD methods. In this paper, we propose a new framework to address this issue. We augment the input image with multiple different views rendered using the depth maps, and cast the conventional single-view RGB-D SOD into a multi-view setting. Since different views captures complementary context of the 3D scene, the accuracy can be significantly improved through multi-view aggregation. We further design a multi-view saliency detection network (MVSalNet), which firstly performs saliency prediction for each view separately and incorporates multi-view outputs through a fusion model to produce final saliency prediction. A dynamic filtering module is also designed to facilitate more effective and flexible feature extraction. Extensive experiments on 6 widely used datasets demonstrate that our approach compares favorably against state-of-the-art approaches.

Keywords: RGB-D Salient Object Detection, multi-view augmentation, multi-view fusion

1 Introduction

RGB-D salient object detection (SOD) aims to identify and segment the most conspicuous objects in the input scene considering both RGB images and the corresponding depth maps. With the rapid development of depth sensors, RGB-D SOD has found wide applications in surveillance [54], autonomous driving [43], and robotics [42], to name a few. Since additional depth information permits comprehensive understanding of the 3D geometry, RGB-D SOD is inherently more superior than its RGB based counterpart in handling challenging scenarios, including background clutter, illumination variation, etc., and therefore has attracted increasingly more attention from the community.

⋆ Corresponding author.

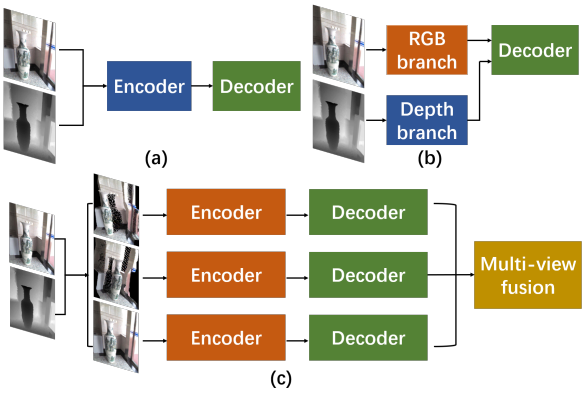

Fig. 1. Framework comparison. (a)(b) Existing RGB-D SOD methods mainly use the input depth map as an additional feature channel. (c) We leverage the 3D geometry of the input depth to perform multi-view saliency detection.

Since the depth map and RGB images are from two different modalities with significant cross-modal gap, it is not a trivial task to perform SOD by simultaneously utilizing the two input data modalities. As such, recent research efforts [5, 30] mainly focus on cross-modal fusion between the input RGB and depth for SOD (cf. Figure 1 (a), (b)). Although significant progress has been achieved, these existing methods mostly use the depth information as an additional input channel to provide low-level cues like edges, contours, and regions, while the essential 3D geometry information are under-explored. This drawback may potentially restrict the merits of existing RGB-D SOD, leading to unsatisfactory performance.

As we humans move freely in the 3D world, we can perceive the scene from different views, allowing more precise foreground detection even at adversarial cases. In fact, the human vision system are also binocular for more effective 3D perception. All these evidences indicate that multi-view perception enabled by 3D geometry can significantly benefit vision tasks.

Motivated by above observations, we propose a new framework to fully explore the geometry information for RGB-D SOD. Instead of using depth map as only low-level cues, we leverage the contained 3D geometry to render the input image under different views, which allows multi-view perception to be mimicked from a single static image. SOD can then be performed for each view independently and the generated single view predictions are eventually fused to produce the final saliency maps (As illustrated in Figure 1 (c)). Since different views may capture different context of input scene and are complementary to each other, the saliency predictions aggregated from multiple views are shown to be more accurate and robust.

We implement the above idea by designing a multi-view saliency detection network (MVSalNet), which contains multiple saliency prediction streams for the

augmented input views, and a multi-view fusion module to incorporate single-view saliency predictions into the final result. To ensure more effective deep feature extraction, we further design a dynamic filtering module (TDF) using transformer networks, which generates position-specific filters according to the input features, leading to more flexible and adaptive convolutions. The entire network can be learned in an end-to-end manner, and compares favorably against existing RGB-D SOD methods.

Our new framework provides an alternative idea for RGB-D SOD. Since single view RGB-D SOD is reformulated as multi-view RGB SOD, the cross-modal gap between RGB image and depth is naturally resolved. Besides, as existing RGB SOD methods can be easily incorporated into our framework for single-view saliency prediction, our method has the potential to benefit from advances in RGB SOD domain.

In summary, the contribution of this paper can be summarized as follows.

- We present a new framework for RGB-D SOD with multi-view augmentation, which can effectively leverage the geometry information carried in input depth maps.
- We design a multi-view saliency prediction network with dynamic filtering modules, which can not only enhance saliency prediction in each single view, but also enables cross-view prediction fusion, yielding more accurate SOD results.

Our method sets new state of the art on 6 benchmark datasets. Extensive evaluation has justified the effectiveness of our contribution.

2 Related Work

2.1 RGB-D Salient Object Detection

Traditional methods are mainly based on hand-crafted features, such as contrast [36], shape [7], compactness [8], background enclosure [16] and so on. As the representation ability of the hand-crafted features is limited, all the above models can not cope with complex scenes. While recently, deep learning-based methods have made significant progress [18, 50, 46] due to the powerful ability in discriminative feature representation. Based on the scope of this paper, we divide existing deep-based models into single-stream models [41, 54] and multi-stream models [15, 48]. The single-stream models directly fuse RGB images and depth maps to send to the network. For example, DANet [52] uses depth-enhanced dual attention to generate contrasted features for the decoder. For the multi-stream models, the frameworks employ parallel networks to extract and fuse multi-modal features with various strategies. For example, Zhang et al. [48] propose an asymmetric two-stream network and design a flow ladder module for RGB stream and a depth attention module for depth stream. Generally speaking, single-stream model is lighter and multi-stream model has better performance.

However, unlike the aforementioned methods in which depth cues are only treated as the direct input of the feature extractor. In this paper, we further exploit the use of depth information. As the depth information contains abundant geometric prior knowledge, we utilize the depth cues to rotate the corresponding RGB images. Then we get multi-view saliency results and fuse them to generate the final output. This results in two major benefits: 1) We generate multi-view RGB images to replace the original depth map, in this way, we explicitly eliminate the modal gap; 2) The noise in low-quality depth map is largely reduced as we use late-fusion [19] to fuse the multi-view saliency results.

2.2 Novel View Synthesis

In some tasks, new views can be synthesized as a data augmentation method. [55] formulates the 3D object detection problem as the detection of rotated bounding boxes in images from bird's eye view generated using the homography. [26] randomly manipulates the camera system, including its focal length, receptive field and location, to generate new training images with geometric shifts. [53] introduces a perspective-aware data augmentation that synthesizes new training examples with more diverse views by perturbing the existing ones in a geometrically consistent manner. Inspired by them, we propose to generate multi-view RGB images in RGB-D SOD.

2.3 Attention Mechanism and Transformer

Fully-convolutional networks [31] are mature architecture for dense prediction, they adopt convolution and subsampling as fundamental elements in order to learn multi-scale features that can leverage an appropriately large context. Furthermore, attention has already proven to be an effective architecture for learning strong models for natural language processing (NLP) [27, 10]. There have been several works that adapt attention mechanisms to computer vision tasks and get competitive results, such as image classification [11], object detection [2], and panoptic segmentation [44]. This is likely because attention can capture long-range associations, which further lead to the trends that combine CNNs with transformers [2, 28]. Notice the advantages of combining the two, we propose to leverage a transformer-based dynamic filtering module to generate adaptive kernels and get more effective features.

3 Method

In this section, we present a new paradigm for RGB-D saliency detection with multi-view augmentation. Figure 2 overviews the pipeline of our method. Given an input image $\mathbf{I}$ and its corresponding depth map $\mathbf{D}$, we first render the RGB image from multiple novel views. Saliency detection is then independently performed under each of the newly rendered views as well as the original input view. Finally, we aggregate all the predicted saliency maps from different views

to produce the output saliency prediction. We implement the above multi-view saliency detection and aggregation procedures through a multi-view saliency detection network (MVSalNet). In the following, we first elaborate on multi-view image synthesis in Section 3.1, and then describe the architecture of our proposed MVSalNet in Section 3.2.

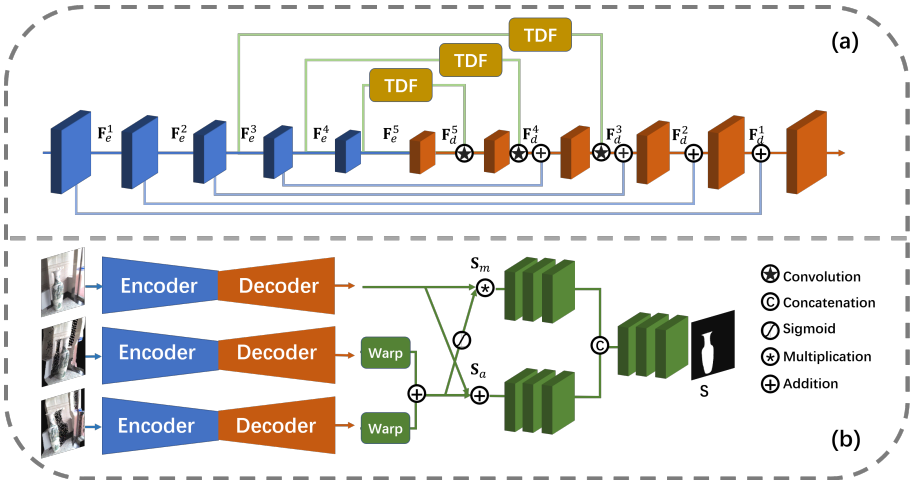

Fig. 2. (a) The details of the encoder-decoder branch in MVSalNet. (b) The overall architecture of MVSalNet. The network has three parallel encoder-decoder branches, and are fed multi-view RGB images, respectively. A multi-view fusion module is added at the end of the network to fuse the multi-view saliency results.

3.1 Multi-View Rendering

As opposed to prior RGB-D SOD methods that mainly use depth as an additional input feature, we propose to explore the 3D geometry information encoded in the depth maps for novel view synthesis, allowing single image RGB-D SOD to be conducted in a multi-view setting. To this end, we develop a multi-view rendering module to efficiently perform multi-view augmentation for the input image. Our basic principle is to reconstruct the 3D point cloud based on the input scene depth, which is then projected to a specific target novel view to render the RGB image.

Technically, given the depth value d of a pixel and its 2D coordinate $\mathbf{p}$ in the input image, its 3D point coordinate $\mathbf{P}$ can be computed. Given the relative motion between the input and a novel target view, we can further transform the 3D point $\mathbf{P}$ to the target view, and then project it onto the target image plane to obtain its corresponding pixel coordinate $\bar{\mathbf{p}}$ in the target image.

The above process establishes a position mapping from each pixel in the input view to its corresponding pixel in the novel target view, based on which we can

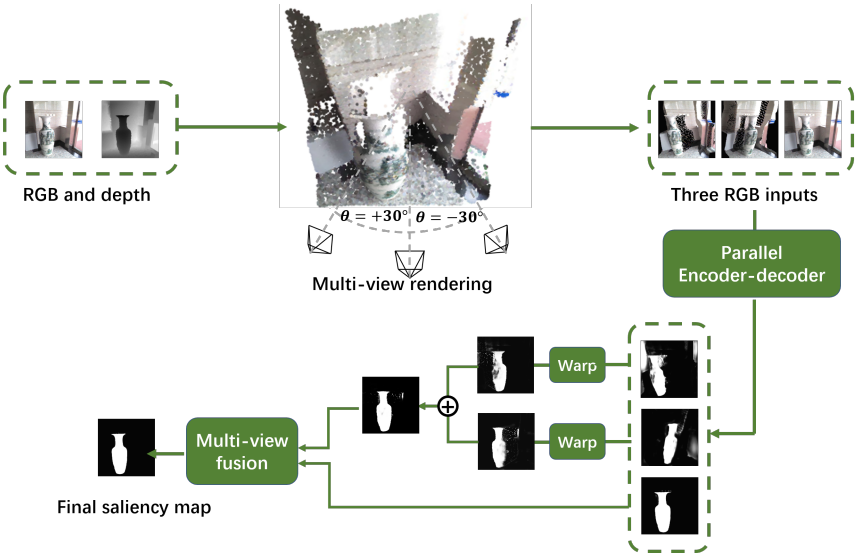

Fig. 3. Implementation of multi-view data augmentation.

synthesize the target view image using the input image, or inversely, warp the predicted saliency map of the target view to the input view. For missing regions in the rendered images or saliency maps caused by occlusion, out-of-view, etc., we fill the missing values with 0.

Considering both efficiency and effectiveness, we augment the input image with two additional novel views. Although augmentation with more novel views may lead to better performance, it will also increase computational overhead. Our preliminary experiment further shows that using fixed relative motions for novel views performs more superior than random generated ones. Therefore, we restrict the rotation of the two novel views on the xy (horizontal) plane in the camera coordinate system. The rotation angles are empirically set to $\pm 30°$ around the z (vertical) axis, respectively. See Figure 3 for an illustration. Since the position of the two novel views are symmetric w.r.t. the original view, the rendered images are complementary in the sense that missing regions in one view will be rendered in the other view (See Figure 4). As a result, the two symmetric views can partially alleviate the impact of occluded or out-of-view regions during novel view rendering.

3.2 Multi-View Saliency Detection Network

We design a multi-view saliency detection network (MVSalNet) with multi-view augmented images as input. As shown in Figure 2 (b), the MVSalNet can be divided into two parts, including the single-view saliency prediction module and multi-view fusion module. Since we augment the input image with two additional views, the single-view saliency prediction module contains three encoder-

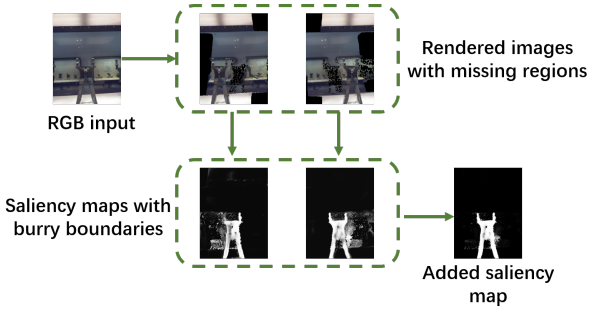

Fig. 4. The complementary property of two symmetric views. Missing regions in one view may be rendered from the other view. The combination of the saliency maps predicted for individual maps can therefore effectively improve detection accuracy.

decoder networks, each of them operating in a specific input view. To further strengthen single-view saliency detection, the encoder and decoder features are skip-connected via the transformer-based dynamic filtering (TDF) module. The predicted saliency maps under different views are further aggregated by the multi-view fusion module to produce the final output.

Single-View Encoder-Decoder. The three encoder-decoder networks under different views share the same architecture with untied network parameters. For the encoder, we adopt the ResNet-50 [20] backbone architecture, which produces a multi-scale feature pyramid denoted as $\{\mathbf{F}_e^i | i = 1, 2, \ldots, 5\}$ with i indicating the resolution index. The feature resolution becomes smaller as the layer goes deeper. The decoder then takes the coarsest-level feature $\mathbf{F}_e^i$ as input and progressively upsamples the intermediate feature maps $\{\mathbf{F}_d^i | i = 5, 4, \ldots, 1\}$ to the original input resolution. Short-cut connections are also added between encoder and decoder features of the same resolutions. Different from existing methods [40] that either use addition or concatenation to combine the corresponding features in the short-cut connections, we design a TDF module (as detailed below) which takes the encoder features $\mathbf{F}_e^3$-$\mathbf{F}_e^5$ and produce three position-specific dynamic filters. The generated filters are then applied to the corresponding decoder features $\mathbf{F}_d^3$-$\mathbf{F}_d^5$, respectively. Each decoder then independently predicts a saliency map for its input view.

Transformer-Based Dynamic Filtering Module. Figure 5 (a) overviews the network architecture of the proposed transformer-based dynamic filtering (TDF) module. For an input feature from the single-view saliency encoder, the TDF module aims to generate a position-specific dynamic filter which can then be applied to the corresponding features in the decoder. Due to its remarkable capabilities in modeling global correlation, we adopt transformer networks for the dynamic filter generation. To this end, we first partition the input encoder feature $\mathbf{F}_e^i$ into 1×1 patches, which are then fed into a linear layer to produce

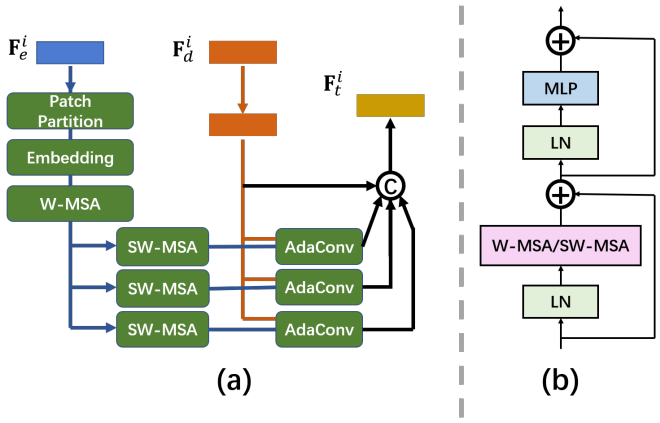

Fig. 5. (a) The structure of TDF. (b) The detail structure of W-MSA and SW-MSA.

a set of 1D feature embeddings corresponding to each location. The embeddings are further processed by a window based multi-head self attention (W-MSA) block [29] followed by three shifted window based MSA (SW-MSA) blocks [29] (cf. Figure 5 (b)), producing three convolutional kernels for each spatial position on the input feature map. The generated convolutional kernels are then applied to the corresponding feature $\mathbf{F}_d^i$ in the decoder through adaptive convolutions [23] with dilation rates of 1, 3, and 5, respectively. The final output $\mathbf{F}_t^i$ of the TDF module can be computed as:

$$\mathbf{F}_t^i = \mathcal{M}(< \mathcal{H}(\mathbf{F}_d^i), \mathcal{H}_a(\mathbf{F}_d^i; \mathbf{K}_1^i, \mathbf{K}_2^i, \mathbf{K}_3^i) >), \tag{1}$$

where $\mathcal{H}$ denotes the standard 3×3 convolution; $\mathcal{H}_a$ denotes adaptive convolution layer using the three generated position specific kernels $\mathbf{K}_1^i$-$\mathbf{K}_3^i$; $< \cdot, \cdot >$ indicates channel-wise concatenation; and $\mathcal{M}$ is a linear transformation layer. As a result, the obtained features are more effective for the decoding of saliency map.

Multi-View Fusion. Figure 2 (b) demonstrates the pipeline of our multi-view fusion module. The single-view encoder-decoders predict the saliency map $\mathbf{S}_0$ for the input image and $\hat{\mathbf{S}}_1$, $\hat{\mathbf{S}}_2$ for the two augmented views. We first warp the saliency predictions for two augmented views to the input view to obtain the saliency maps $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively. Considering their complimentary property, we add the warped augmented view together as $\mathbf{S}_{1,2} = \mathbf{S}_1 + \mathbf{S}_2$ to tackle occluded or out-of-view regions with missing values. To achieve more effective multi-view fusion, we adopt both element-wise multiplication and addition to combine the current and augmented view saliency maps:

$$\begin{aligned} \mathbf{S}_m &= \mathbf{S}_0 \odot \sigma(\mathbf{S}_{1,2}), \\ \mathbf{S}_a &= \mathbf{S}_0 + \mathbf{S}_{1,2}, \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ denotes sigmoid function. Intuitively, $\mathbf{S}_m$ is able to suppress false-positive background noises while $\mathbf{S}_a$ allows to identify false-negative foreground

regions. We then concatenate both $\mathbf{S}_m$ and $\mathbf{S}_a$ along the channel dimension and send their concatenation to an additional convolution layer to generate the final saliency map $\mathbf{S}$ of the input image.

Loss Function. For the loss function, we directly use the binary cross entropy (BCE) loss with the hybrid enhanced loss (HEL) in [35].

BCE loss is the common loss in SOD task, the main form is as follows:

$$L_b = -\sum[\mathbf{G}\log(\mathbf{S}) + (1-\mathbf{G})\log(1-\mathbf{S})], \tag{3}$$

where $\mathbf{S}$ and $\mathbf{G}$ respectively represent the prediction and the corresponding ground truth. the loss of each supervised saliency map is expressed as follows:

$$L = L_b + L_h, \tag{4}$$

where L is the loss of each supervised saliency map, L_h is HEL.

So the total loss of the network can be calculated by the following formula:

$$L_t = L(\mathbf{S}) + \alpha{\cdot}L(\mathbf{S_0}) + \beta{\cdot}L(\mathbf{S}_{1,2}), \tag{5}$$

where L_t is the total loss of the network, α and β are the weight coefficients, we set $\alpha = \beta = 0.25$ in this paper.

4 Experiments

In this section, we conduct extensive experiments to verify the effectiveness of our method. First, we compare our model with other methods. Then we perform a series of ablation studies to evaluate each component of our framework.

4.1 Datasets and Evaluation Metrics

Datasets. We perform our experiments on six widely used RGB-D datasets for fair comparisons. LFSD [25] contains 100 image pairs. NJUD [24] contains 1985 image pairs. NLPR [36] contains 1000 image pairs. RGBD135 [6] contains 135 image pairs. STEREO [34] contains 1000 stereoscopic image pairs. DUTRGBD [38] 1200 image pairs. To guarantee fair comparisons, we follow the setting of [38]. On the DUTRGBD dataset, we choose the same 800 samples for training and 400 images for testing. For the other datasets, we follow the data partition of [3] to use 1485 samples from NJUD and 700 samples from NLPR to train and the remaining samples are used to test.

Evaluation metrics. To comprehensively and fairly evaluate various methods, we employ five widely used metrics for evaluating, including F-measure [1], weighted F-measure [33], MAE [37], S-measure [12], and E-measure [13]. F-measure [1] reflects the performance of the binary predictions under different

	Metric	TANet[4]	A2dele[39]	HDFNet[35]	JL-DCF[17]	UCNet[47]	DANet[52]	BTSNet[49]	Ours
DUTRGBD [38]	F_{max}	0.862	0.906	0.930	0.924	0.882	0.918	0.929	0.935
	F_{ada}	0.815	0.891	0.885	0.883	0.856	0.888	0.906	0.914
	F_{β}^{ω}	0.764	0.865	0.864	0.863	0.822	0.860	0.872	0.893
	MAE	0.067	0.042	0.041	0.043	0.056	0.043	0.039	0.034
	S_m	0.853	0.884	0.907	0.905	0.863	0.899	0.903	0.915
	E_m	0.901	0.929	0.938	0.938	0.906	0.937	0.942	0.951

Table 1. Quantitative RGB-D SOD results on DUTRGBD dataset. The best results are highlighted in red.

thresholds. Weighted F-measure is proposed to improved the existing metric F-measure, it defines a weighted precision and a weighted recall. MAE measures the average of the per-pixel absolute difference between the saliency maps and the ground truth. S-measure can evaluate the structural similarities. E-measure can jointly utilize image-level statistics and local pixel-level statistics for evaluating the binary saliency map.

4.2 Implementation Details

Parameter setting. Three encoders of the proposed model are based on ResNet-50 [20], and only the convolutional layers in the corresponding classification networks are retained. During the training phase, we use the weight parameters pretrained on the ImageNet [9] to initialize the encoders.

Training setting. During the training stage, we apply random horizontal flipping, random rotating as data augmentation for RGB images and depth images to improve generalization and avoid overfitting. And we employ random color jittering for RGB images. We use the momentum SGD optimizer with a weight decay of 5e-4, an initial learning rate of 5e-3, and a momentum of 0.9. Besides, we apply the CosineAnnealing strategy [32] with the minimum learning rate of 0. The input images are resized to 320×320. We train the model for 40 epochs on a NVIDIA GTX 1080 Ti GPU with a batch size of 4.

Testing details. During the testing stage, we resize RGB and depth images to 320×320. The final prediction is rescaled to the original size for evaluation.

4.3 Comparisons

To demonstrate the effectiveness of the proposed method, we compare it with 14 state-of-the-art (SOTA) methods, including TANet [4], D3Net [14], A2dele [39], AFNet [45], CoNet [22], CPFP [51], JL-DCF [17], PCF [3], UCNet [47], HDFNet [35], DCF [21], BBSNet [15], DANet[52], BTSNet [49]. Quantitative results on the DUTRGBD dataset are shown in Table 1, while those on the rest five datasets are shown in Table 2. Our methods consistently outperforms all the other SOTAs across all the datasets in terms of different metrics.

	Metric	TANet [4]	D3Net [14]	A2dele [39]	AFNet [45]	CoNet [22]	CPFP [51]	JL-DCF [17]	PCF [3]	UCNet [47]	HDFNet [35]	DCF [21]	BBSNet [15]	BTSNet [49]	Ours
LFSD[25]	F_{max}	0.827	0.849	-	0.780	0.874	0.850	0.872	-	0.871	0.872	0.861	0.879	0.849	0.880
	F_{ada}	0.794	0.801	-	0.742	0.835	0.813	0.830	-	0.844	0.833	0.815	0.850	0.823	0.856
	F_β^ω	0.719	0.756	-	0.671	0.802	0.775	0.792	-	0.813	0.789	0.776	0.815	0.770	0.819
	MAE	0.111	0.099	-	0.133	0.077	0.088	0.082	-	0.072	0.088	0.087	0.073	0.098	0.072
	S_m	0.801	0.832	-	0.738	0.856	0.828	0.847	-	0.851	0.841	0.828	0.860	0.829	0.856
	E_m	0.851	0.860	-	0.810	0.892	0.867	0.885	-	0.896	0.885	0.865	0.902	0.874	0.906
NJUD[24]	F_{max}	0.888	0.903	0.888	0.804	0.900	0.890	0.914	0.887	0.906	0.921	0.920	0.922	0.927	0.922
	F_{ada}	0.844	0.840	0.873	0.768	0.780	0.837	0.881	0.844	0.885	0.887	0.898	0.894	0.901	0.902
	F_β^ω	0.805	0.833	0.844	0.696	0.848	0.828	0.866	0.803	0.867	0.877	0.886	0.879	0.884	0.886
	MAE	0.061	0.051	0.051	0.100	0.047	0.053	0.042	0.059	0.043	0.037	0.036	0.038	0.035	0.035
	S_m	0.878	0.895	0.868	0.772	0.895	0.878	0.902	0.877	0.894	0.909	0.908	0.915	0.918	0.910
	E_m	0.909	0.901	0.916	0.847	0.924	0.900	0.935	0.909	0.932	0.930	0.936	0.933	0.942	0.939
NLPR[36]	F_{max}	0.876	0.904	0.895	0.816	0.895	0.883	0.924	0.864	0.911	0.926	0.914	0.921	0.912	0.929
	F_{ada}	0.796	0.834	0.878	0.747	0.844	0.818	0.868	0.795	0.885	0.887	0.887	0.882	0.874	0.901
	F_β^ω	0.780	0.826	0.859	0.693	0.838	0.807	0.873	0.762	0.872	0.881	0.881	0.875	0.869	0.895
	MAE	0.041	0.034	0.028	0.058	0.031	0.038	0.023	0.044	0.026	0.024	0.022	0.024	0.027	0.021
	S_m	0.886	0.906	0.895	0.799	0.904	0.884	0.921	0.873	0.912	0.924	0.920	0.924	0.920	0.927
	E_m	0.916	0.934	0.945	0.884	0.933	0.920	0.953	0.916	0.952	0.955	0.958	0.952	0.949	0.959
RGBD135 [6]	F_{max}	0.853	0.917	0.893	0.775	0.908	0.882	0.931	-	0.931	0.932	0.903	0.934	0.929	0.934
	F_{ada}	0.795	0.876	0.868	0.730	0.866	0.829	0.899	-	0.916	0.908	0.870	0.901	0.899	0.908
	F_β^ω	0.740	0.831	0.838	0.641	0.845	0.787	0.892	-	0.901	0.897	0.844	0.879	0.873	0.903
	MAE	0.046	0.030	0.029	0.068	0.030	0.038	0.021	-	0.019	0.020	0.026	0.023	0.023	0.019
	S_m	0.858	0.904	0.883	0.770	0.906	0.872	0.929	-	0.927	0.929	0.897	0.926	0.917	0.931
	E_m	0.919	0.956	0.919	0.874	0.944	0.927	0.967	-	0.974	0.969	0.947	0.961	0.961	0.971
STEREO[34]	F_{max}	0.878	0.897	-	0.848	0.908	0.889	0.915	-	0.903	0.908	0.909	0.907	0.905	0.920
	F_{ada}	0.835	0.833	-	0.807	0.879	0.830	0.858	-	0.875	0.862	0.875	0.874	0.870	0.898
	F_β^ω	0.787	0.815	-	0.752	0.864	0.817	0.850	-	0.857	0.846	0.863	0.847	0.848	0.879
	MAE	0.060	0.054	-	0.075	0.038	0.051	0.041	-	0.041	0.044	0.040	0.043	0.044	0.035
	S_m	0.871	0.891	-	0.825	0.902	0.879	0.901	-	0.895	0.896	0.897	0.901	0.899	0.911
	E_m	0.916	0.911	-	0.887	0.939	0.907	0.932	-	0.939	0.928	0.937	0.933	0.932	0.946

Table 2. Results (↑:F_{max}, F_{ada}, F_β^ω, S_m, and E_m; ↓:MAE) of different RGB-D SOD methods across five datasets. The best results are highlighted in red.

Figure 6 shows sampled visualization results under challenging scenarios, including cluttered background (Row 3, 4, 10), complex objects (Row 2, 7, 8, 9), low-quality depth map (Row 1, 6), and small objects with misleading depth map (Row 5).

4.4 Ablation Study

In this section, we perform a series of ablation studies on the NLPR dataset [36] to further investigate the relative importance and specific contribution of each component in the proposed framework using as test dataset.

multi-view augmentation. To validate the effectiveness of our multi-view augmentation, we conduct several experiments. Results are shown in Table 3. "ED" means that we only use one encoder-decoder branch with RGB input. "EDaug" means that we use one encoder-decoder branch and add the RGB images from novel views to the training set to get extra training data. "3ED" means that we use three encoder-decoder branches with the same RGB input.

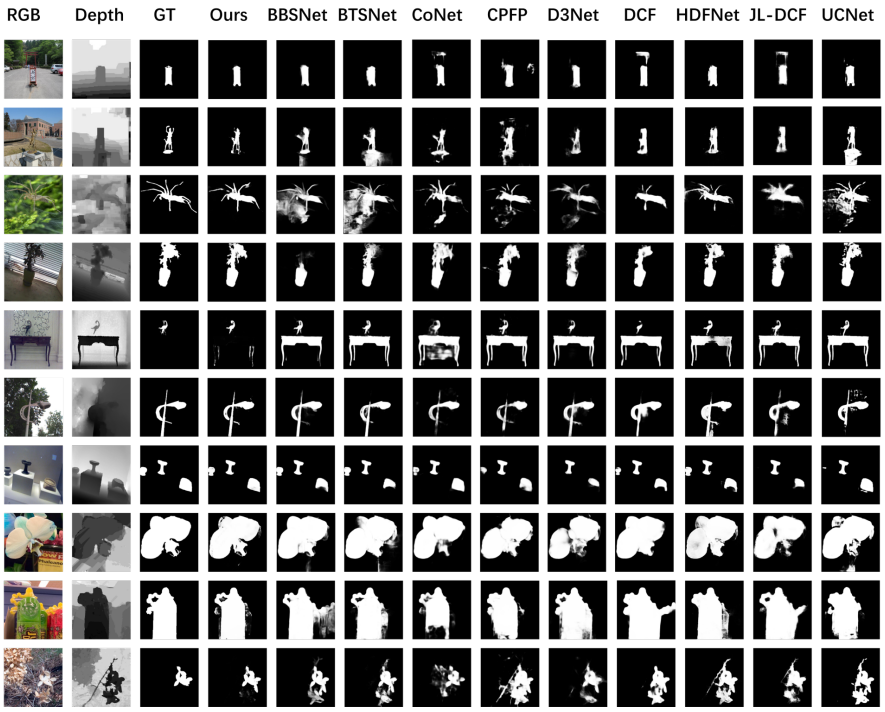

Fig. 6. Visual comparisons with SOTA RGB-D SOD models.

"Ours-1view" means that we only generate one novel view, the two images pass through two parallel encoder-decoders. "Ours-depth" means that we use our proposed network and change the input of the original encoder-decoder branch to the concatenation of the original RGB image and depth map. It shows that the multi-view augmentation can effectively improve performance when used as extra input or just as extra data. Notice that "Ours-depth" uses more data and gets lower performance compared to "Ours", suggesting that our method can make full use of depth information, while a vanilla single-view encoder-decoder may fail to address the modal gap between RGB and depth. "Ours-15°", "Ours-45°", and "Ours-60°" mean that the rotation angles are set to $\pm 15°$, $\pm 45°$, and $\pm 60°$, respectively. Our preliminary experiments show that rotation angles larger than 60° will lead to degraded results. "Ours-random" means that we randomly choose the rotation angle between 0° and 60° in each side. "Ours-4views" means that we add two more views at $\pm 60°$ and they share weight with $\pm 30°$ branches, respectively. Compared with "Ours-1view" and "Ours-4views", we can see that "Ours" reach a good balance between efficiency and effectiveness.

Dynamic Filtering module. Ablations of the dynamic filtering module are reported in Table 4. "SK" means using normal skip connection without TDF. "no dilation" means that we only use one SW-MSA head with dilation rate of

Model	F_{max}	F_{ada}	F_β^ω	MAE	S_m	E_m
ED	0.919	0.860	0.855	0.029	0.910	0.939
EDaug	0.919	0.874	0.868	0.027	0.916	0.948
3ED	0.928	0.891	0.882	0.023	0.920	0.954
Ours-1view	0.922	0.882	0.876	0.024	0.917	0.951
Ours-depth	0.922	0.895	0.889	0.023	0.923	0.958
Ours-15°	0.929	0.894	0.889	0.021	0.924	0.957
Ours-45°	0.929	0.893	0.887	0.022	0.924	0.958
Ours-60°	0.929	0.889	0.884	0.022	0.923	0.955
Ours-random	0.928	0.894	0.890	0.021	0.925	0.958
Ours-4views	0.929	0.899	0.888	0.021	0.927	0.962
Ours	0.929	0.901	0.895	0.021	0.927	0.959

Table 3. Ablation on the multi-view augmentation. The best results are highlighted in red.

Model	F_{max}	F_{ada}	F_β^ω	MAE	S_m	E_m
SK	0.922	0.886	0.878	0.025	0.918	0.952
no dilation	0.926	0.892	0.885	0.022	0.921	0.957
dilation embed	0.927	0.896	0.886	0.022	0.924	0.957
DDPM	0.929	0.896	0.889	0.022	0.922	0.956
Ours	0.929	0.901	0.895	0.021	0.927	0.959

Table 4. Ablation on dynamic filtering module. The best results are highlighted in red.

Model	F_{max}	F_{ada}	F_β^ω	MAE	S_m	E_m
add	0.923	0.883	0.881	0.023	0.920	0.951
supervise two	0.927	0.896	0.890	0.022	0.923	0.958
Ours	0.929	0.901	0.895	0.021	0.927	0.959

Table 5. Ablation on multi-view fusion. The best results are highlighted in red.

Model	F_{max}	F_{ada}	F_β^ω	MAE	S_m	E_m
$\alpha = \beta = 0$	0.925	0.893	0.883	0.024	0.920	0.955
$\alpha = \beta = 0.5$	0.928	0.901	0.895	0.021	0.926	0.960
$\alpha = \beta = 1$	0.928	0.894	0.886	0.022	0.923	0.957
Ours	0.929	0.901	0.895	0.021	0.927	0.959

Table 6. Ablation on loss function. The best results are highlighted in red.

1. "dilation embed" means that we use one SW-MSA head but embed features using dilated convolution with dilation rates of 1, 3, 5 in parallel and concatenate the outputs. "DDPM" means that we replace TDF module in our model with the CNN dynamic filtering branch in [35]. It shows that our proposed TDF module outperforms the counterparts. Besides, "DDPM" has 389M parameters, "Ours" has 83M parameters, the number of parameters is reduced by 78.6%.

Multi-view Fusion. Ablations of multi-view fusion are reported in Table 5. Among them, "add" means directly adding the three saliency maps in original

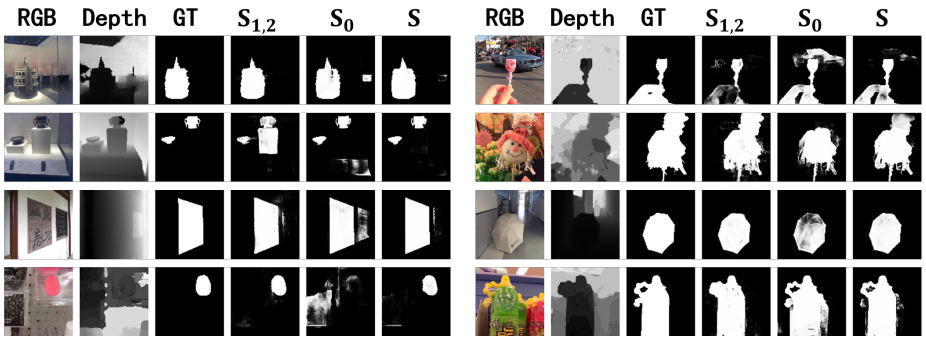

Fig. 7. Visual comparisons for showing the benefits of multi-view fusion. GT, $S_{1,2}$, and S_0 denotes ground truth of saliency map, the added saliency map from augmented views, and the saliency map from original view, respectively.

view. "supervise two" denotes that the saliency maps of the two synthesized RGB images are supervised in original view respectively. It shows that our light-weight module can learn the complementary property between the saliency maps and generate accurate final result. In Figure 7, we can see that S_0 mainly focuses on texture information, while $S_{1,2}$ captures abundant spatial structure information. The final result S is generated by fusing S_0 and $S_{1,2}$.

Loss Function. Ablations of the weight coefficients α and β in loss function are reported in Table 6. Experiments show that our model is robust to the hyper parameters α and β.

5 Conclusion

In this paper, we propose a new RGB-D salient object detection (SOD) framework to take full advantages of 3D geometry information contained in depth maps. Instead of using input depth maps as low-level cues, we render the input image from multiple different views and formulate SOD from a single static images to a multi-view setting. We further design a multi-view salient detection network (MVSalNet), which performs SOD independently for each individual view and fuses the output from multiple views to obtain the final prediction. The proposed method outperforms state-of-the-art RGB-D SOD approaches on 6 benchmark datasets with a considerable margin, which demonstrates the effectiveness of our contributions.

Acknowledgements This paper is supported by National Natural Science Foundation of China (61725202, U1903215, 61906031, 61829102), and Fundamental Research Funds for Central Universities (DUT21RC(3)025).

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1597–1604. IEEE (2009)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
3. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for rgb-d salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3051–3060 (2018)
4. Chen, H., Li, Y.: Three-stream attention-aware network for rgb-d salient object detection. IEEE Transactions on Image Processing **28**(6), 2825–2835 (2019)
5. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. Pattern Recognition **86**, 376–385 (2019)
6. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method. In: Proceedings of international conference on internet multimedia computing and service. pp. 23–27 (2014)
7. Ciptadi, A., Hermans, T., Rehg, J.M.: An in depth view of saliency. Georgia Institute of Technology (2013)
8. Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C.: Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Processing Letters **23**(6), 819–823 (2016)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557 (2017)
13. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018)
14. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on neural networks and learning systems **32**(5), 2075–2089 (2020)
15. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In: European Conference on Computer Vision. pp. 275–292. Springer (2020)
16. Feng, D., Barnes, N., You, S., McCarthy, C.: Local background enclosure for rgb-d salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2343–2350 (2016)

17. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q.: Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3052–3062 (2020)
18. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q., Shen, J., Zhu, C.: Siamese network for rgb-d salient object detection and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
19. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE transactions on cybernetics **48**(11), 3171–3183 (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., et al.: Calibrated rgb-d salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9471–9481 (2021)
22. Ji, W., Li, J., Zhang, M., Piao, Y., Lu, H.: Accurate rgb-d salient object detection via collaborative learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. pp. 52–69. Springer (2020)
23. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. Advances in neural information processing systems **29**, 667–675 (2016)
24. Ju, R., Liu, Y., Ren, T., Ge, L., Wu, G.: Depth-aware salient object detection using anisotropic center-surround difference. Signal Processing: Image Communication **38**, 115–126 (2015)
25. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2806–2813 (2014)
26. Lian, Q., Ye, B., Xu, R., Yao, W., Zhang, T.: Geometry-aware data augmentation for monocular 3d object detection. arXiv preprint arXiv:2104.05858 (2021)
27. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
28. Liu, Y., Sun, G., Qiu, Y., Zhang, L., Chhatkuli, A., Van Gool, L.: Transformer in convolutional neural networks. arXiv preprint arXiv:2106.03180 (2021)
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
30. Liu, Z., Zhang, W., Zhao, P.: A cross-modal adaptive gated fusion generative adversarial network for rgb-d salient object detection. Neurocomputing **387**, 210–220 (2020)
31. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
32. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
33. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 248–255 (2014)

34. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 454–461. IEEE (2012)
35. Pang, Y., Zhang, L., Zhao, X., Lu, H.: Hierarchical dynamic filtering network for rgb-d salient object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 235–252. Springer (2020)
36. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: a benchmark and algorithms. In: European conference on computer vision. pp. 92–109. Springer (2014)
37. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 733–740. IEEE (2012)
38. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7254–7263 (2019)
39. Piao, Y., Rong, Z., Zhang, M., Ren, W., Lu, H.: A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9060–9069 (2020)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
41. Shigematsu, R., Feng, D., You, S., Barnes, N.: Learning rgb-d salient object detection using background enclosure, depth contrast, and top-down features. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 2749–2757 (2017)
42. Skoczeń, M., Ochman, M., Spyra, K., Nikodem, M., Krata, D., Panek, M., Pawłowski, A.: Obstacle detection system for agricultural mobile robot application using rgb-d cameras. Sensors **21**(16), 5292 (2021)
43. Wan, T., Du, S., Cui, W., Yao, R., Ge, Y., Li, C., Gao, Y., Zheng, N.: Rgb-d point cloud registration based on salient object detection. IEEE transactions on neural networks and learning systems (2021)
44. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. pp. 108–126. Springer (2020)
45. Wang, N., Gong, X.: Adaptive fusion for rgb-d salient object detection. IEEE Access **7**, 55277–55284 (2019)
46. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F., Aliakbarian, S., Barnes, N.: Uncertainty inspired rgb-d saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
47. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8582–8591 (2020)
48. Zhang, M., Fei, S.X., Liu, J., Xu, S., Piao, Y., Lu, H.: Asymmetric two-stream architecture for accurate rgb-d saliency detection. In: European Conference on Computer Vision. pp. 374–390. Springer (2020)
49. Zhang, W., Jiang, Y., Fu, K., Zhao, Q.: Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)

50. Zhang, Z., Lin, Z., Xu, J., Jin, W.D., Lu, S.P., Fan, D.P.: Bilateral attention network for rgb-d salient object detection. IEEE Transactions on Image Processing **30**, 1949–1961 (2021)
51. Zhao, J.X., Cao, Y., Fan, D.P., Cheng, M.M., Li, X.Y., Zhang, L.: Contrast prior and fluid pyramid integration for rgbd salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3927–3936 (2019)
52. Zhao, X., Zhang, L., Pang, Y., Lu, H., Zhang, L.: A single stream network for robust and real-time rgb-d salient object detection. In: European Conference on Computer Vision. pp. 646–662. Springer (2020)
53. Zhao, Y., Kong, S., Fowlkes, C.: Camera pose matters: Improving depth prediction by mitigating pose distribution bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15759–15768 (2021)
54. Zhu, C., Cai, X., Huang, K., Li, T.H., Li, G.: Pdnet: Prior-model guided depth-enhanced network for salient object detection. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 199–204. IEEE (2019)
55. Zhu, M., Zhang, S., Zhong, Y., Lu, P., Peng, H., Lenneman, J.: Monocular 3d vehicle detection using uncalibrated traffic cameras through homography. arXiv preprint arXiv:2103.15293 (2021)