Is Depth Really Necessary for Salient Object Detection?

Jiawei Zhao[1,†], Yifan Zhao[1,†], Jia Li[1,2,3*], Xiaowu Chen[1]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University, Beijing, 100191, China
[2]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, 100191, China
[3]Peng Cheng Laboratory, Shenzhen, 518066, China

ABSTRACT

Salient object detection (SOD) is a crucial and preliminary task for many computer vision applications, which have made progress with deep CNNs. Most of the existing methods mainly rely on the RGB information to distinguish the salient objects, which faces difficulties in some complex scenarios. To solve this, many recent RGBD-based networks are proposed by adopting the depth map as an independent input and fuse the features with RGB information. Taking the advantages of RGB and RGBD methods, we propose a novel depth-aware salient object detection framework, which has following superior designs: 1) It does not rely on depth data in the testing phase. 2) It comprehensively optimizes SOD features with multi-level depth-aware regularizations. 3) The depth information also serves as error-weighted map to correct the segmentation process. With these insightful designs combined, we make the first attempt in realizing an unified depth-aware framework with only RGB information as input for inference, which not only surpasses the state-of-the-art performance on five public RGB SOD benchmarks, but also surpasses the RGBD-based methods on five benchmarks by a large margin, while adopting less information and implementation light-weighted.

CCS CONCEPTS

• **Computing methodologies → Interest point and salient region detections**.

KEYWORDS

salient object detection, depth awareness, RGBD

ACM Reference Format:
Jiawei Zhao[1,†], Yifan Zhao[1,†], Jia Li[1,2,3*], Xiaowu Chen[1]. 2020. Is Depth Really Necessary for Salient Object Detection?. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413855

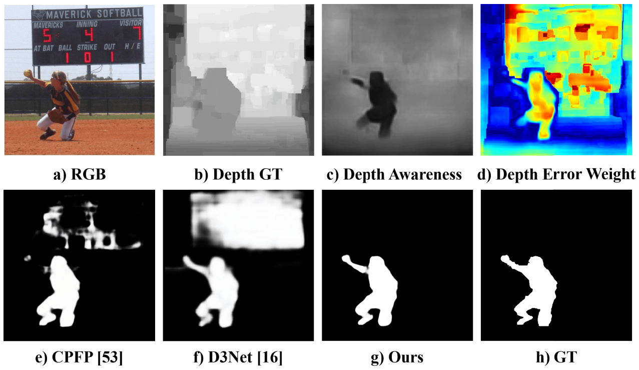
a) RGB
b) Depth GT
c) Depth Awareness
d) Depth Error Weight
e) CPFP [53]
f) D3Net [16]
g) Ours
h) GT

Figure 1: Motivation of our depth-aware salient object detection. b): captured depth groundtruth. c): predicted depth awareness by DASNet. d): depth-aware error weights for salient correction. e) and f) are generated by two RGBD SOD models.

1 INTRODUCTION

Salient object detection (SOD) aims to detect and segment objects that attract human attention most visually. With the proposals of large datasets [23, 28, 34, 35, 44, 48] and deep learning techniques [20, 32], recent works have made significant progress in accurately segmenting salient objects, which can serve as an important prerequisite for a wide range of computer vision tasks, such as semantic segmentation [26], visual tracking [21], and image retrieval [39].

Recent years have witnessed significant progress in the field of salient object detection. Previous works [8, 24, 25, 31, 37, 41, 47, 48, 54] take only the RGB information as inputs, which is relatively lightweight and can be easily trained end-to-end. For example, Wu *et al.* [47] propose a coarse-to-fine feature aggregation framework to generate saliency maps. However, the reasoning of salient regions can not be well solved when there exist multiple contrasting region proposals or ambiguous object contours. Therefore, the depth information can be a complementary guidance to deduct the overlapping and viewpoint issues, which can be beneficial to salient object detection.

Combing the RGB information with the auxiliary depth inputs, recent research efforts [19, 36, 53] have verified its effectiveness in improving the object segmentation process. These methods usually introduce an additional depth stream to encode depth map and then fuse the RGB stream and depth stream to deduct the salient objects. For example, Han *et al.* [19] propose a two-stream network to extract RGB features and depth features, and then fuse them

[†]Jiawei Zhao and Yifan Zhao contribute equally to this work.
[*]Jia Li is the corresponding author (E-mail: jiali@buaa.edu.cn).
Website: https://cvteam.net/.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413855

with a combination layer. Piao *et al.* [36] propose a two-stream network and fuse paired multi-level side-out features to refine the final saliency results. The main drawbacks of RGBD-based methods are twofold. On the one hand, the additional depth branch introduces heavy computation costs compared to the methods with bare RGB inputs. On the other hand, the object segmentation process heavily relies on the acquisition of depth maps, which are usually unavailable in some extreme occasions or realistic industrial applications. Keeping these cues in our mind, a natural concern arises: is depth information really necessary for salient object detection and what roles should depth play in salient object detection?

Taking the essence and discarding the dregs of RGB and RGBD methods, we set out to create a unified framework, which only takes the depth information as supervision in the training stage. Hence the network can take only the RGB images as inputs, and meanwhile is aware of depth prior knowledge with the learnt network parameters. That is to say, we make use of depth information to regularize the learning process of salient object detection (See Fig. 1). First, we force the feature maps in different levels of network to be aware of depth information. This can be conducted in a multi-task learning trend when learning the object segmentation and estimating the depth map simultaneously. The estimated depth awareness map can be found in Fig. 1 c). Although the estimated depth map is not highly accurate as captured one (in Fig. 1 b)), but focuses on more contrastive depth regions, which are desirable for the segmentation process. Second, the estimated depth awareness can also be considered as an indicator to find the most ambiguous regions. We calculate the logarithmic error map of the estimation and ground truth depth to generate an adaptive weight map in Fig. 1 d). The network is further forced to pay more attention to pixels with higher error-weighted responses, hence some semantic confusions can be improved. Comparing to other state-of-the-art RGBD-based models [16, 53] in Fig. 1 e) and f), the proposed approach can better tackle the salient confusions while generating clear object boundaries.

In this paper, we make three insightful designs to construct our framework in Fig. 2, which make full use of training data from multiple sources. *i.e.*, data from RGB source and RGBD source can be separately fed into this framework with different learning constraints to promote the final performance. To achieve this framework, we first propose a depth awareness module, to regularize the features in different levels of the network stage while learning the object segmentation in the meantime. This forces the segmentation features to be aware of constrastive object in the depth of field. Second, we propose a generalized channel-aware fusion model (CAF) to aggregate the features from top to bottom levels in these two relevant branches. Then the final depth features and segmentation features are fused with the same CAF module in this coarse-to-fine scheme. Last but not least, we utilize a depth error-weighted map to emphasize the saliency ambiguous regions, *i.e.*, objects salient in images but not in depth, or vice versa. These regions are attached with more attention in the overall learning procedure to alleviate the object confusions and generate clear object boundaries. Experimental evidences demonstrate the effectiveness in promoting RGBD salient object detection with only RGB inputs and the potential in promoting RGB tasks with auxiliary training depth.

Contributions of this paper are summarized as follows: 1) We first set out a novel setting to use depth data as training priors to facilitate the salient object detection and propose a unified framework to solve this important problem. 2) We propose a channel-aware fusion model (CAF) to comprehensively fuse multi-level features, which can retain rich details and pay more attention to the significant features. 3) We propose a novel joint depth awareness module to facilitate the understanding of saliency and design a depth-aware error loss to mine ambiguous pixels. 4) Experimental evidences demonstrate that the proposed model achieves the state-of-the-art performance both on five RGBD benchmarks and five RGB benchmarks.

2 RELATED WORK

RGB-based Salient Object Detection. Early traditional RGB SOD methods mainly rely on hand-crafted cues such as color constrant [8], texture [48] and local/global contrast [25]. Borji *et al.* [1] comprehensively review these methods for details with both deep learning and conventional techniques. Recently CNN-based RGB SOD methods have achieved impressive improvements over traditional methods [8, 25, 48]. Most of them follow an end-to-end architecture as shown in Fig. 3 a). Liu *et al.* [31] utilize pixel-wise contextual attention to selectively attend to global and local context information. Wu *et al.* [47] propose a coarse to fine aggregation framework, which discards low-level features to reduce the complexity. Zhao *et al.* [54] propose a pyramid feature attention network, which adopts channel-wise attention and spatial attention to focus more on valuable features. Su *et al.* [41] propose a boundary-aware network to fuse the boundary and interior features with a compensation mechanism and an adaptive manner. Qin *et al.* [37] design a hybrid loss to focus on the boundary quality of salient objects. Wei *et al.* [24] propose a cross-feature module to fuse features of different levels.

RGBD-based Salient Object Detection. Although existing RGB methods have achieved very high performance, they might fail when dealing with complex scenarios, *e.g.*, low contrast, occlusions. It is shown that depth is an important and effective cue for saliency detection [12] especially in these complex scenarios. Existing RGB-D SOD methods mainly rely on extracting salient features from RGB image and depth map respectively, and then fuse them in the early or late network stages. Peng *et al.* [35] directly concatenate RGB-D pairs as 4-channel inputs to predict saliency maps. Han *et al.* [19] propose a two-stream network to extract RGB features and depth features, and then fuse them with a combination layer. Chen *et al.* [2] propose a progressive fusion strategy in a coarse-to-fine manner. Zhao *et al.* [54] propose a fluid pyramid integration strategy to make full use of depth enhanced features. Piao *et al.* citepiao2019depth develop a two-stream network and fuse paired multi-level side-out features to refine the final salient object detection.

Single Image Depth Estimation. Monocular depth estimation can be divided into three categories according to the input: monocular video [43, 46], stereo image pairs [18, 42] and single image [13, 14, 17, 27, 50], in which taking single image as input is the hardest case because there is no geometric information in only a single image. Thanks to the powerful deep networks like VGG [40] and ResNet [20], single image depth estimation has been

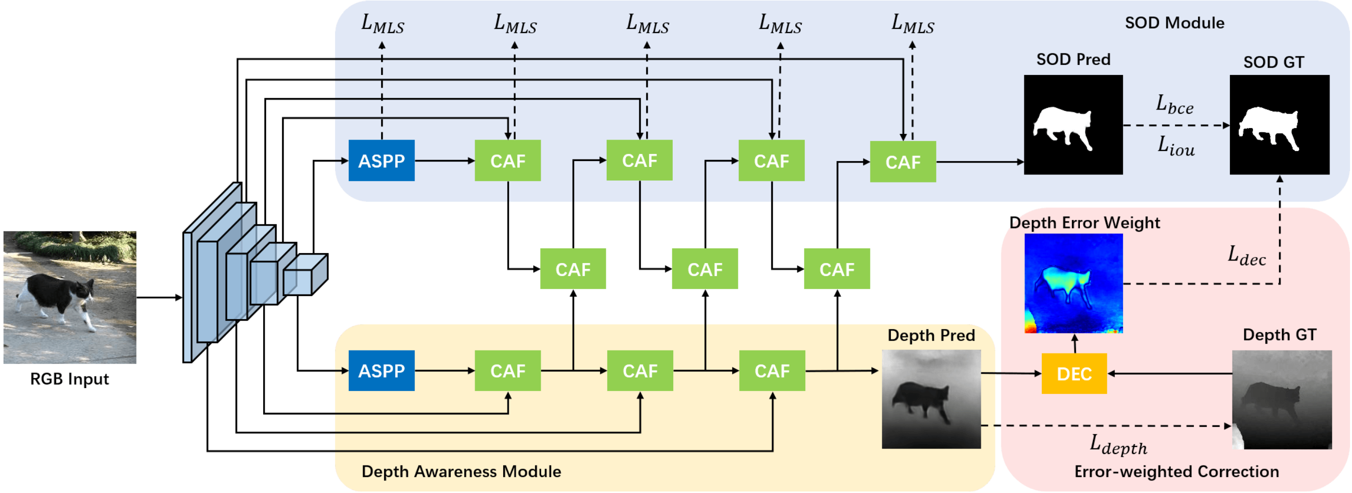

Figure 2: The overall architecture of our model. Our depth-awareness SOD framework is mainly composed of three parts, *i.e.*, a salient object detection module, a depth awareness module and an error-weighted correction. ASPP denotes atrous spatial pyramid pooling. CAF denotes the proposed channel-aware fusion module. DEC denotes the proposed depth error-weighted correction. The dashed line denotes supervision.

boosted to a new accuracy level. Eigen *et al.* [13, 14] propose the first CNN-based framework for single image depth estimation, which applies a stage-wisely multi-scale network to refine depth estimation. Laina *et al.* [27] introduce a fully convolutional architecture and design reverse Huber loss to smoothness effect of L2 norm. Fu *et al.* [17] propose a spacing-increasing discretization strategy to discretize depth and recast depth estimation as an ordinal regression problem. Yin *et al.* [50] propose a global geometric constraint to improve the depth estimation accuracy. As an important cue in many vision tasks, there are many works utilize multi-task learning to joint depth estimation and other pixel-level vision tasks, such as semantic segmentation [33], surface normal [50].

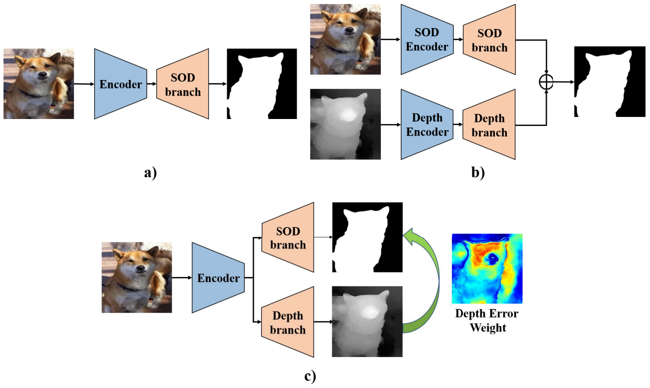

Figure 3: Different types of SOD architecture. a) : Typical RGB-based SOD network architecture. b): Typical RGBD-based SOD network architecture. c): Proposed Depth-awareness SOD network architecture.

3 METHODOLOGY

3.1 Overview

Depth-Awareness SOD Network. In this section, we present a novel joint Depth-Awareness SOD Network (DASNet) for RGBD-based and RGB-based salient object detection tasks, which is mainly composed of three modules, *i.e.*, the SOD module, the depth awareness module and the depth error-weighted correction (see Fig. 2). The first two modules share similar structures but focus on different tasks, which are supervised by saliency maps and depth maps respectively. The SOD module and the depth awareness module utilize our proposed channel-aware fusion model (CAF) to fuse high-level and low-level features. Taking these two branches into combination, we finally refine the saliency results by the proposed depth error-weighted constraint, which could mine hard pixels with the supervision of depth maps.

Relations and Discussions. Our intuitive idea comes from the RGB and RGBD salient object detection tasks, which is shown in Fig. 3. The conventional RGB SOD in Fig. 3 a) takes the original image as input with a encoder-decoder framework. With the depth as auxiliary input in Fig. 3 b), the overall framework requires two independent encoders to extract the depth and RGB features separately, which main computation costs are usually lied on. Moreover, the depth and RGB encoders are separately trained and the relationships between these multi-modal data are not fully explored.

Taking only RGB inputs as well as saving the computation costs, the depth-aware salient object detection in Fig. 3 c) provides us a new perspective to utilize the depth data in this segmentation task. In the testing phase, the network only takes the RGB as input and the object segmentation results are regularized by the depth-awareness constraints in the training phase. In this manner, the network not only builds an explicit relationship between depth and SOD, but also saves the additional costs in feature extraction.

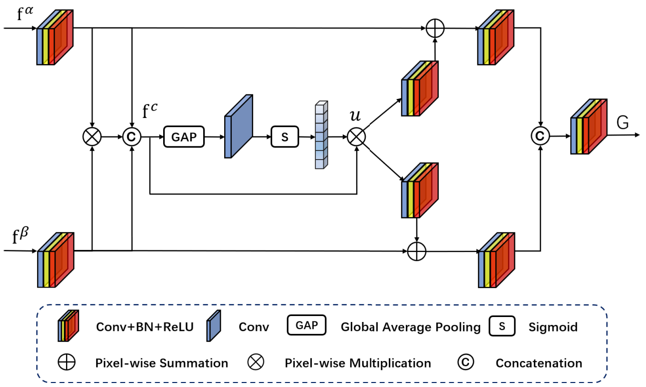

Figure 4: The proposed channel-aware fusion module. Blocks denote basic convolutional units and G is the fused output.

3.2 Channel-Aware Fusion Module

The crucial problem in salient object detection is to select the most discriminative features and pass them in the coarse-to-fine scheme. However, aggregating features from different levels in a encoder-decoder fashion usually leads to missing details or introduces ambiguous features, which both make the network fail to optimize. Notably, this phenomenon appears more frequently when it comes to aggregating features from different domains. Therefore, a selective feature fusion strategy is in high demand, especially for RGBD salient object understanding.

Toward this end, we propose a novel Channel-Aware Fusion module (CAF), which adaptively selects the discriminative features for object understanding. Instead of using different specific structures for different aggregation strategies in previous works [7, 36, 41], we advocate using a generalized module to fuse any common types of features, *e.g.*, features from different levels and features from different sources.

The proposed CAF has some meaningful designs, which are illustrated in Fig. 4. First, given two types of source feature $\mathbf{f}^{\alpha}, \mathbf{f}^{\beta} \in \mathbb{R}^{W' \times H' \times C'}$, we use pixel-wise multiplication to enhance the common pixels in feature maps, while alleviate the ambiguous ones. The enhanced features are then concatenated with the transformed features with a lightweight encoder $\xi(\cdot)$. It can be formally represented as:

$$\mathbf{f}^{c} = \xi_{\alpha}(\mathbf{f}^{\alpha}) © \xi_{\beta}(\mathbf{f}^{\beta}) © (\xi_{\alpha}(\mathbf{f}^{\alpha}) \otimes \xi_{\beta}(\mathbf{f}^{\beta})), \tag{1}$$

where © and ⊗ denote the feature concatenation operation and pixel-wise multiplication respectively. Each encoder $\xi_{\{\alpha,\beta\}}$ is typically composed of a 3×3 convolutional layer followed by a Batch Normalization and a ReLU activation. Specially, when aggregating the multi-level features, the features $\mathbf{f}^{\alpha}$ and $\mathbf{f}^{\beta}$ are first upsampled to the same scale, which is omitted for better view in Fig. 4.

After obtaining rich feature $\mathbf{f}^{c} \in \mathbb{R}^{W' \times H' \times 3C'}$ by (1), the second main concern is how to select the most relevant features that are highly-responded in the segmentation target. Inspired by channel-attention mechanism [5, 22], we thus propose to use global features for a contextual understanding in the attention weights. The $\mathbf{f}^{c}$ are then squeezed with a global average pooling, followed by a sigmoid normalization σ, and transformed as the vector shape to align the dimensions with feature channels. This serialized operation has the form:

$$\mathbf{g} = \frac{1}{W' \times H'} \sum_{i=1}^{W'} \sum_{j=1}^{H'} \mathbf{f}^{c}_{i,j}, \tag{2}$$

$$\mathbf{u}_{i,j} = \mathbf{f}^{c}_{i,j} \otimes \sigma(\varphi_c(\mathbf{g}_{i,j})). \tag{3}$$

φ is a linear transformation to reorganize the pooling features and $\mathbf{u}$ denotes the learnt attention weighted features. Therefore features relevant to the salient target could be prominent in each group of source features $\mathbf{f}^{\alpha}$ and $\mathbf{f}^{\beta}$. This can be achieved by a channel-aware attention mechanism:

$$\mathcal{G} = \tau_g(\tau_{v1}(\xi_{\alpha}(f^{\alpha}) \oplus \xi_{u1}(\mathbf{u})) © \tau_{v2}(\xi_{\beta}(f^{\beta}) \oplus \xi_{u2}(\mathbf{u}))), \tag{4}$$

where $\xi_{\{u1,u2\}}$ denotes the typical decoder and $\tau_{\{v1,v2,g\}}$ denotes the typical decoder with dimensional reductions as original input. Hence the relevant features to target object can be enhanced in the final output $\mathcal{G}$. In addition, to implement the whole framework in a lightweight trend, the channel dimension C' is empirically set as 64 to achieve the state-of-the-art performance.

3.3 Depth-awareness Constraint

What roles does depth play in salient object detection? To answer this aforementioned question, in this paper, we propose an innovative depth-awareness constraint from two complementary aspects, *i.e.*, multi-level depth awareness and depth error-weighted correction. These two aspects work collaboratively to regularize the salient features being aware of contrastive depth regions and contextual salient confusions, which facilitates the segmentation process in different learning stages.

Multi-level Depth Awareness. As mentioned in Section 3.2, the key issue in salient object detection lies on the utilization of multi-level features in different network stages. Besides the aggregation strategy, the other exploitation is to regularize the features focusing on meaningful regions, which would provide useful contextual information before aggregation. Taking the advantages of depth information and the hierarchical network architecture, we force the segmentation features to focus on depth regions in different network learning stages, which is elaborated in Fig. 2. This means in each network learning stages, the features should be aware of the object information as well as the contrastive depth regions. We use an additional depth branch to regress the ground-truth depth.

With this collaborative learning of SOD and depth regression, we further fuse these two modules to refine the salient object (see Fig. 2), which builds strong correlations between these two different types of features. Notably, this refinement strategy can also be well handled by our proposed CAF, with the same segmentation supervision at multiple levels. As a result, the salient features stand as a predominant place in the final optimization and the depth map becomes a leading guidance.

Depth Error-weighted Correction. To make a thorough exploitation of depth information, we further propose a depth error-weighted correction (DEC) which aims to regularize hard pixels with higher weights if the predicted depth make mistakes. As it stands, the network itself naturally tends to be highly responded to the salient regions and then form a holistic salient object. However, this would guide the predicted depth features focusing on salient regions, and cause a severe misalignment between the predicted depth and ground truth data. Remarkably, the error regions where the predicted depth make mistakes are usually the semantic ambiguous regions, which we need to pay more attention to the learning process.

In order to solve this misalignment as well as to exploit it, we thus introduce a logarithmic depth error weight. Let p^d and y^d be the predicted depth and groundtruth depth respectively, the error weight $\mathbf{e}_{ij}$ of each pixel has the form:

$$\mathbf{e}_{ij} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w}(\log p_{ij}^d - \log y_{ij}^d)}{\sum_{i=1}^{h}\sum_{j=1}^{w}\max(\log p^d - \log y^d)}, \tag{5}$$

where w and h are the width and height of the error window, which aims to represent the error of central pixel with the mean value of a local region. The detailed ablations to decide w and h can be found in Tab. 5. In this way, the ambiguous pixels are treated with more attention in the early learning phase. With the optimization goes through, the regularized features become depth-aware and errors are progressively corrected. This learning progress is exhibited in Fig. 5, where the highly-responded regions in the error map shrink along with the learning stage. This verifies that the final optimized features are aware of depth information and better at handling semantic confusions.

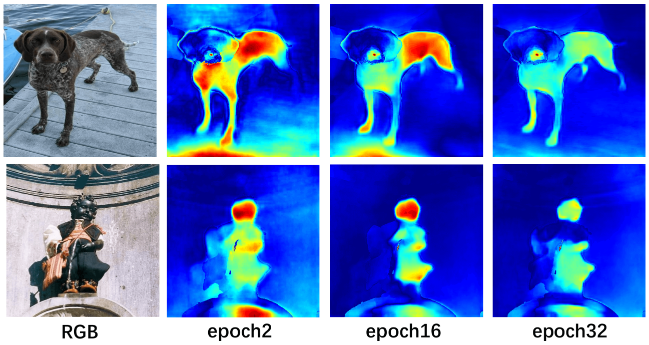

Figure 5: Qualitative visualization of depth error weights during the training stage, with epoch 2, 16, and 32.

3.4 Learning Objective

Our overall learning objective is composed of three modules, as in Fig. 2, the SOD module, the depth awareness module and the error-weighted correction. Let $p^s, y^s \in \mathbb{R}^{W\times H\times 1}$ be the predicted salient mask and corresponding groudtruth, the SOD module is supervised with the BCE loss:

$$\mathcal{L}_{bce} = -\sum_{i=1}^{H}\sum_{j=1}^{W} y_{ij}^s \log(p_{ij}^s). \tag{6}$$

However, the BCE loss usually leads to noisy predictions which does not form a holistic object. To make the salient object with clear boundaries, we adopt a IoU (Intersection over Union) loss [24, 37] as the auxiliary loss:

$$\mathcal{L}_{iou} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(y_{ij}^s \times p_{ij}^s) + 1}{\sum_{i=1}^{H}\sum_{j=1}^{W}(y_{ij}^s + p_{ij}^s - y_{ij}^s \times p_{ij}^s) + 1}. \tag{7}$$

For the depth awareness module, we adopt the log mean square error (logMSE) for supervision [13, 14] to generate smooth depth map, and meanwhile providing the error weights $\mathbf{e}$:

$$\mathcal{L}_{depth} = \frac{1}{W\times H}\sum_{i=1}^{H}\sum_{j=1}^{W} ||\log y_{ij}^d - \log p_{ij}^d||_2^2. \tag{8}$$

For the error-weighted correction module, we adopt a error-weighted BCE loss to attach more importance to wrongly-predicted pixels:

$$\mathcal{L}_{dec} = \frac{-\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{e}_{ij}\times y_{ij}^s\log(p_{ij}^s)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{e}_{ij}}. \tag{9}$$

This error loss $\mathcal{L}_{dec}$ adopts the same supervision as the SOD module with a binary segmentation mask. To implement the multi-level supervision in a unified framework, the overall loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{depth} + \sum_{i=1}^{S}\lambda_i(\mathcal{L}_{bce} + \mathcal{L}_{iou} + \mathcal{L}_{dec}), \tag{10}$$

where λ_i denotes the weight of different level loss and S is set as 5 with five stages in ResNet. Here we follow GCPANet [7] and set λ as [1, 0.8, 0.6, 0.4, 0.2].

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

RGBD-based SOD Datasets. To evaluate the RGBD performance of the proposed approach, we conduct experiments on five benchmarks [9, 23, 34, 35, 55], including NJUD [23] with 1,985 images captured by Fuji W3 stereo camera, NLPR [35] with 1,000 images captured by Kinect, STEREO [34] with 1,000 images collected in the Internet, DES [9] with 135 images captured by Kinect, SSD [55] with 80 images picked up from stereo movies. Following [36, 53], We split 1,500 samples from NJUD and 700 samples from NLPR for training, the rest images in these two datasets and the other three datasets are used for testing.

RGB-based SOD Datasets. To verify the effectiveness for RGB datasets, we adopt five RGB benchmarks [28, 29, 44, 48, 49], including DUTS [44] with 15,572 images, ECSSD [48] with 1,000 images, DUT-OMRON [49] with 5,168 images, PASCAL-S [29] with 850 images, HKU-IS [28] with 4,447 images. DUTS is currently the largest SOD dataset, following [44], we split 10,553 images (DUT-TR) from DUTS for training and 5,019 images (DUT-TE) from DUTS for testing, the other four datasets are also used for testing.

Evaluation Metrics. To quantitatively evaluate the performance of our approach and state-of-the-art methods, we adopt 4 commonly used metrics: max F-measure (F_β^{max}), mean F-measure (F_β^{mean}), mean absolute error (MAE) and structure similarity measure (S_α) [15] on both RGB-based methods and RGBD-based methods.

We use F_β to measure both Precision and Recall comprehensively. F_β is computed based on Precision and Recall pairs as follows:

$$F_\beta = \frac{(1+\beta^2)\times Precision\times Recall}{\beta^2\times Precision + Recall}, \tag{11}$$

where we set β^2=0.3 to emphasize more on Precision than Recall, and compute F_β^{max}, F_β^{mean} using different thresholds as in [1].

Table 1: Performance comparison with 9 state-of-the-art RGBD-based SOD methods on five benchmarks. Smaller MAE, larger F_β^{max}, F_β^{mean} and S_α indicates better performance. The best results are highlighted in bold.

methods	NJUD-TE				NLPR-TE				STEREO				DES				SSD			
	F_β^{max}	F_β^{mean}	MAE	S_α	F_β^{max}	F_β^{mean}	MAE	S_α	F_β^{max}	F_β^{mean}	MAE	S_α	F_β^{max}	F_β^{mean}	MAE	S_α	F_β^{max}	F_β^{mean}	MAE	S_α
DF [38]	.804	.744	.141	.763	.778	.682	.085	.802	.757	.616	.141	.757	.766	.566	.093	.752	.735	.709	.142	.747
AFNet [45]	.775	.764	.100	.772	.771	.755	.058	.799	.823	.806	.075	.825	.728	.713	.068	.770	.687	.672	.118	.714
CTMF [19]	.845	.788	.085	.849	.825	.723	.056	.860	.831	.786	.086	.848	.844	.765	.055	.863	.729	.709	.099	.776
MMCI [4]	.852	.813	.079	.858	.815	.729	.059	.856	.863	.812	.068	.873	.822	.750	.065	.848	.781	.748	.082	.813
PCF [2]	.872	.844	.059	.877	.841	.794	.044	.874	.860	.845	.064	.875	.804	.763	.049	.842	.807	.786	.062	.841
TANet [3]	.874	.844	.060	.878	.863	.796	.041	.886	.861	.828	.060	.871	.827	.795	.046	.858	.810	.767	.063	.839
CPFP[53]	.876	.850	.053	.879	.869	.840	.036	.888	.874	.842	.051	.879	.838	.815	.038	.872	.766	.747	.082	.807
DMRA [36]	.886	.872	.051	.886	.879	.855	.031	.899	.868	.847	.066	.835	.888	.857	.030	.900	.844	.821	.058	.857
D3Net [16]	.889	.860	.051	.895	.885	.853	.030	.904	.881	.844	.054	.904	.885	.859	.030	.904	.847	.818	.058	.866
Ours	**.911**	**.894**	**.042**	**.902**	**.929**	**.907**	**.021**	**.929**	**.915**	**.894**	**.037**	**.910**	**.928**	**.892**	**.023**	**.908**	**.881**	**.857**	**.042**	**.885**

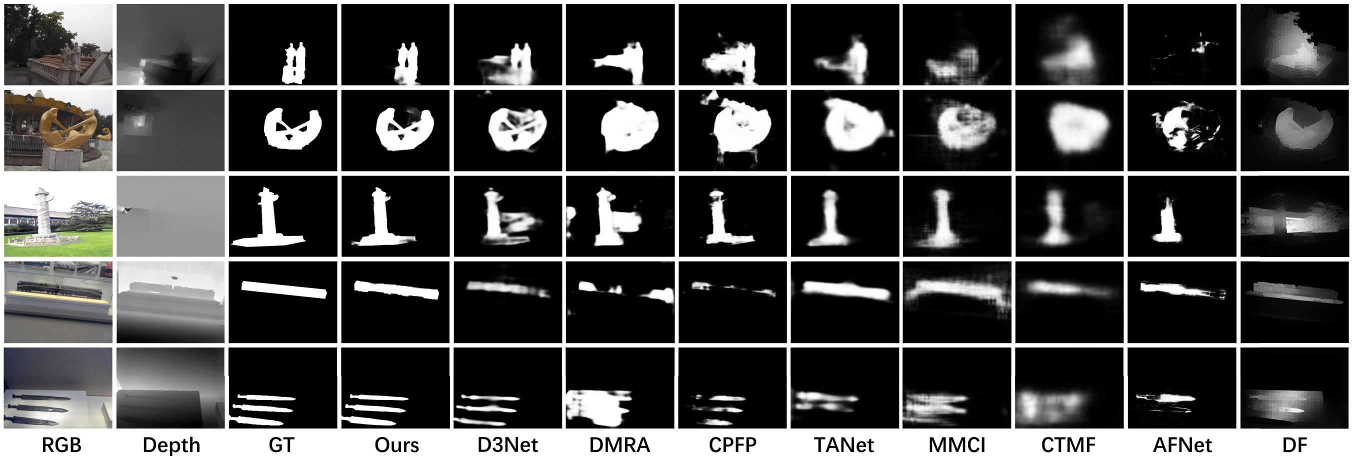

Figure 6: Qualitative comparison of the state-of-the-art RGBD-based methods and our approach. Obviously, saliency maps produced by our model are clearer and more accurate than others in various challenging scenarios.

We use S_α to measure structure similarity for a more comprehensive evaluation. S_α combines the region-aware (S_r) and object-aware (S_o) structural similarity as follows:

$$S_\alpha = \alpha \times S_o + (1 - \alpha) \times S_r, \tag{12}$$

where we set α=0.5 as suggested in [15].

4.2 Implementation Details

We adopt ResNet-50 [20] pre-trained on ImageNet [10] as our backbone. The atrous rate of ASPP follows the prior work [6], which is set as (6, 12, 18). In the training stage, we resize each image to 352×352 and adopt horizontal flip, random crop and multi-scale resize as data augmentation. We use SGD optimizer with the batch size=32 for 32 epochs. Inspired by [7, 24], we adopt warm-up and linear decay strategies to adjust the learning rate with the maximum learning rate 0.005 for ResNet-50 backbone and 0.05 for other parts. We set momentum and decay rate to 0.9 and 5e-4, respectively. It only takes us 1 hour for RGBD-based task and 3 hours for RGB-based task to train a model on one NVIDIA 1080Ti GPU.

For the RGBD-based salient object detection, we utilize both RGB images and depth maps from training sets to train our model. During the testing stage, we only need RGB images as inputs to predict saliency maps on RGBD test sets. For the RGB-based salient object detection task, we first estimate depth maps for DUT-TR by pre-trained VNLNet [50] directly, which works well in single image depth estimation task. Then we utilize both DUT-TR and its corresponding predicted depth maps to train our model. During the inference stage, we only need RGB images as inputs to predict saliency maps on RGB test sets.

4.3 Comparisons with the state-of-the-art

RGBD-based SOD Benchmark. As shown in Tab. 1, we compare our model denoted as DASNet with 9 state-of-the art methods, including DF [38], AFNet [45], CTMF [19], MMCI [4], PCF [2], TANet [3], CPFP [53], DMRA [36], D3Net [16]. For fair comparisons, we obtain the saliency maps from the reported results. Our proposed approach surpasses 9 state-of-the-art RGBD-based saliency detection methods on five benchmarks. As shown in Tab. 1, it is obvious that our method achieves a new performance leader-board with no depth image as input, which puts our model in inferior places for comparison. Especially for the F_β^{max} and F_β^{mean} metric, our model outperforms over 3%, which means our method has a good capability to utilize depth information for more precise saliency maps.

In Fig. 6, we exhibit the saliency maps predicted by our model and other approaches. Among all the methods, our model performs

Table 2: Performance comparison with 10 state-of-the-art RGB-based SOD methods on five benchmarks. Smaller MAE, larger F_β^{max},F_β^{mean} and S_α correspond to better performance. The best results are highlighted in bold.

methods	ECSSD				DUT-TE				DUT-OMRON				HKU-IS				PASCAL-S			
	F_β^{max}	F_β^{mean}	MAE	S_α	F_β^{max}	F_β^{mean}	MAE	S_α	F_β^{max}	F_β^{mean}	MAE	S_α	F_β^{max}	F_β^{mean}	MAE	S_α	F_β^{max}	F_β^{mean}	MAE	S_α
BMPM [51]	.929	.894	.045	.911	.851	.762	.049	.861	.774	.698	.064	.808	.921	.875	.039	.905	.862	.803	.073	.840
PAGR [52]	.927	.894	.061	.889	.854	.784	.056	.838	.771	.711	.071	.775	.918	.886	.048	.887	.854	.803	.094	.815
R3Net [11]	.929	.883	.051	.910	.829	.716	.067	.837	.793	.690	.067	.819	.910	.853	.047	.894	.837	.775	.101	.809
PiCA-R [31]	.935	.901	.046	.917	.860	.759	.051	.869	.803	.717	.065	.832	.919	.880	.043	.905	.867	.800	.077	.852
BANet [41]	.939	.917	.041	.924	.872	.829	.040	.879	.782	.750	.061	.832	.923	.893	.037	.913	.847	.839	.079	.852
PoolNet [30]	.944	.915	.039	.921	.880	.809	.040	.883	.808	.747	.055	.833	.933	.899	.032	.916	.869	.822	.074	.845
BASNet [37]	.943	.880	.037	.916	.859	.791	.048	.866	.805	.756	.056	.836	.928	.895	.032	.909	.857	.775	.078	.832
CPD-R [47]	.939	.917	.037	.918	.865	.805	.043	.869	.797	.747	.056	.825	.925	.891	.034	.905	.864	.824	.072	.842
F3Net [24]	.945	.925	.033	.924	.890	.840	.035	.888	.813	.766	.053	.838	.937	.910	.028	.917	.880	.840	**.064**	.855
GCPANet [7]	.948	.919	.035	**.927**	.888	.817	.040	.891	.812	.748	.056	.839	.938	.898	.031	.920	.876	.836	**.064**	**.861**
Ours	**.950**	**.932**	**.032**	**.927**	**.896**	**.853**	**.034**	**.894**	**.827**	**.783**	**.050**	**.845**	**.942**	**.917**	**.027**	**.922**	**.885**	**.849**	**.064**	.860

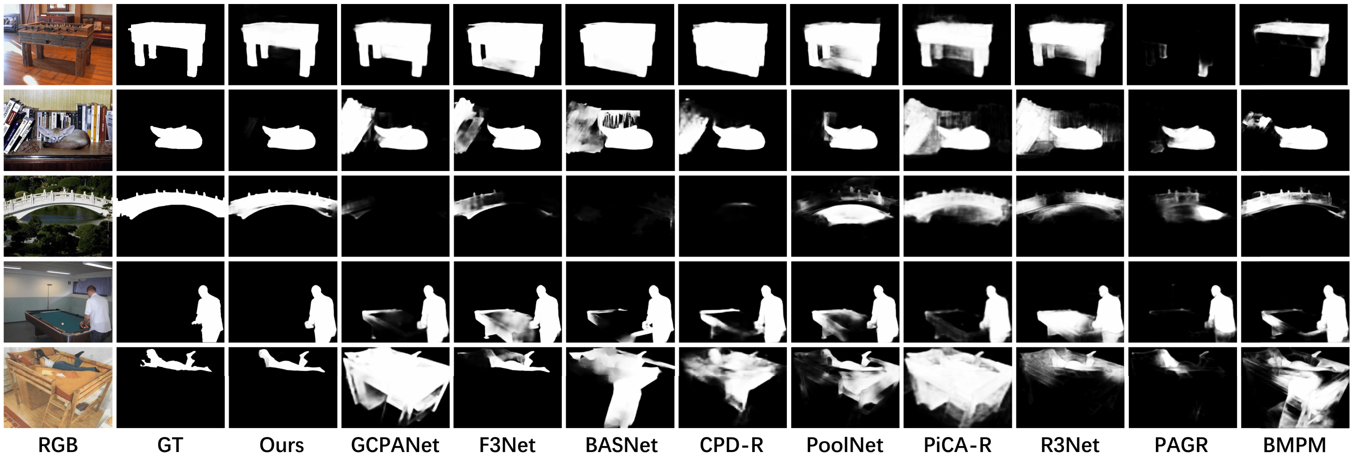

Figure 7: Qualitative comparison of the state-of-the-art RGB-based methods and our approach. Obviously, saliency maps produced by our model are clearer and more accurate than others in various challenging scenarios.

best both on completeness and clarity. In the first, second, and third rows, our method could obtain more accurate and clearer saliency maps than others with ambiguous depth cues. In forth and fifth rows, our method could obtain more complete results than others. Our proposed framework could utilize depth cues much better in various challenging scenarios. Besides, the object boundaries predicted by our model are clearer and sharper than others.

RGB-based SOD Benchmark. As shown in Tab. 2, we compare our proposed DASNet with 10 state-of-the-art methods, *i.e.*, BMPM [51], PAGR [52], R3Net [11], PiCANet [31], PoolNet [30], BANet [41], CPDNet [47], BASNet [37], F3Net [24], GCPANet [7]. As shown in Tab. 2, we can see our proposed DSANet still outperforms other methods and ranks first on all datasets and almost all metrics. However, this performance is achieved with only estimated depth maps as training priors. we believe that with the captured real data, the final performance would be improved steadily, which is vailidated on the RGBD benchmarks.

As shown in Fig. 7, comparing with visual results of different methods, our approach shows an advantage in completeness and clarity. In first and second rows, our method could distinguish foreground and background and obtain more accurate results than other methods in complex scenarios with similar foreground and background. In third row, our method could obtain more complete results in complex scenarios with low contrast, while other methods might fail to detect salient objects in the same scenarios. In forth and fifth rows, our method can provide accurate object localization when salient objects touching image boundaries. Besides, the object boundaries predicted by our model are clearer and sharper than others.

4.4 Performance Analysis

To investigate the effectiveness of each key component in our proposed model, we first conduct a thorough ablation study and then measure the computation complexity for the state-of-the-art models to show its superiority. Finally an experiment for finding hyperparameters can be found in Tab. 5.

Channel-Aware Fusion. To evaluate the effectiveness of our feature fusion module, we reconstruct our model with different ablation factors. Tab. 3 shows the ablations on NJUD-TE dataset. In the first row, we first build our model with widely-used lateral connections between different levels of features, and then fuse them by pixel-wise summation as our baseline. In the second row, we replace the fusion strategy aforementioned with proposed CAF.

Table 3: Ablation study for different components. BCE, IoU, DEC are different loss functions mentioned above. CAF denotes the proposed channel aware fusion module. DAM denotes the depth awareness module. MLS represents multi-level supervision.

BCE	CAF	IoU	DAM	DEC	MLS	NJUD-TE	
						F_β^{mean}	MAE
✓						.838	.058
✓	✓					.853	.056
✓	✓		✓			.857	.051
✓	✓		✓	✓		.871	.048
✓	✓	✓				.875	.047
✓	✓	✓	✓			.880	.045
✓	✓	✓	✓	✓		.886	.043
✓	✓	✓	✓	✓	✓	.894	.042

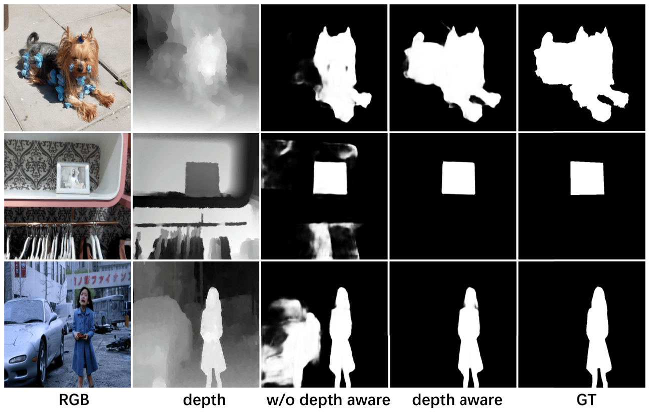

Figure 8: Qualitative results on RGBD datasets. The third column without depth awareness is hard to distinguish complex scenarios with similar foreground and background, while our model in the forth column shows better performance.

This more effective fusion strategy can improve F_β^{mean} of baseline from 0.838 to 0.853.

Depth-awareness Constraint. Then we test our proposed DAM and DEC on the baseline using only BCE loss , and both BCE and IoU loss respectively. Comparing with model using CAF and only BCE loss, our proposed DAM and DEC can improve F_β^{mean} 1.8% in total. Compared with the baseline using CAF and both BCE loss and IoU loss, our proposed DAM and DEC can improve F_β^{mean} from 0.875 to 0.886 and MAE from 0.047 to 0.043. At last, we add multi-level supervision to refine our results. As shown in Tab. 3, all components contribute to the performance improvement, which demonstrates the necessity of each component of our proposed model to obtain the best saliency detection results. Qualitative results can be found in Fig. 8. In the third column, our model without the DAM and DEC would be confused in regions with similar foreground and background. With DAM and DEC, our model could distinguish these confusing features and generate more accurate and clearer saliency maps.

Table 4: Complexity comparison with RGB-based models and RGBD-based models. Models ranking the first and second place are viewed in bold and underlined.

	Methods	Platform	Params(M)	MAdds(G)
RGB&RGBD	Ours	pytorch	**36.68**	<u>11.57</u>
RGB	GCPANet [7]	pytorch	67.06	26.61
	BASNet [37]	pytorch	87.06	97.51
	CPD-R [47]	pytorch	<u>47.85</u>	**7.19**
	BANet [41]	caffe	55.90	35.83
RGBD	CPFP [53]	caffe	72.94	21.25
	DMRA [36]	pytorch	59.66	113.09

Table 5: Error correction results on NLPR-TE with different window sizes.

	1×1	3×3	5×5	7×7	15×15	31×31
F_β^{max}	.924	**.929**	.925	**.929**	.926	.927
F_β^{mean}	.895	.904	.898	**.907**	.904	.897
MAE	.024	**.021**	.022	**.021**	.023	.022
S_α	.924	.928	.926	**.929**	.926	.925

Computational Efficiency. Tab. 4 shows the parameters and computational cost measured by multiply-adds (MAdds) of our proposed model and other open-sourced RGB-based models and RGBD-based models. Our model could achieve obvious higher performance in a light-weight fashion. Notably, CPD-R [47] discards features of two shallower layers to improve the computation efficiency, but sacrifices the accuracy and clarity of results. For fair comparisons, we obtain the deployment codes released by authors and evaluate them with the same configuration.

Hyper-paramters. To evaluate the effectiveness as well as to find the adequate window sizes in (5), we tune the $w \times h$ to be different sizes and choose 7×7 to achieve the best performance. This means that the error weight should be locally aware thus to generate clear object details. This also indicates that amplifying the local receptive field of error-weighted correction module in an adequate range is effective to reach higher scores.

5 CONCLUSIONS

In this paper, we rethink the problem of depth in the field of salient object detection and propose a new perspective of containing the depth constraints in learning process, rather than using the captured depth as inputs. To make a deeper exploitation of depth information, we develop a multi-level depth awareness constraint and a depth error-weighted loss to alleviate the salient confusions. These advanced designs endow our model lightweight and be free of depth input. Experimental results reveal that with only RGB inputs, the proposed network not only surpasses the state-of-the-art RGBD methods by a large margin but well demonstrates its effectiveness in RGB application scenarios.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (No. 61922006 and No. 61532003), and the Beijing Nova Program (No. Z181100006218063).

REFERENCES

[1] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. 2015. Salient object detection: A benchmark. *IEEE transactions on image processing* 24, 12 (2015), 5706–5722.

[2] Hao Chen and Youfu Li. 2018. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3051–3060.

[3] Hao Chen and Youfu Li. 2019. Three-stream attention-aware network for RGB-D salient object detection. *IEEE Transactions on Image Processing* 28, 6 (2019), 2825–2835.

[4] Hao Chen, Youfu Li, and Dan Su. 2019. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognition* 86 (2019), 376–385.

[5] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 5659–5667.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[7] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. 2020. Global Context-Aware Progressive Aggregation Network for Salient Object Detection. In *AAAI Conference on Artificial Intelligence (AAAI).* 10599–10606.

[8] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. 2014. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2014), 569–582.

[9] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. 2014. Depth enhanced saliency detection method. In *Proceedings of international conference on internet multimedia computing and service.* 23–27.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 248–255.

[11] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. 2018. R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence.* AAAI Press, 684–690.

[12] Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. 2013. Depth really Matters: Improving Visual Salient Region Detection with Depth. In *Proceedings of the British Machine Vision Conference (BMVC).*

[13] David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision.* 2650–2658.

[14] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems.* 2366–2374.

[15] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision.* 4548–4557.

[16] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. 2020. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. *IEEE TNNLS* (2020).

[17] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2002–2011.

[18] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision.* Springer, 740–756.

[19] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. 2017. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE transactions on cybernetics* 48, 11 (2017), 3171–3183.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[21] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. 2015. Online tracking by learning discriminative saliency map with convolutional neural network. In *International conference on machine learning.* 597–606.

[22] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 7132–7141.

[23] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. 2014. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE international conference on image processing (ICIP).* IEEE, 1115–1119.

[24] Qingming Huang Jun Wei, Shuhui Wang. 2020. F3Net: Fusion, Feedback and Focus for Salient Object Detection. In *AAAI Conference on Artificial Intelligence (AAAI).*

[25] Dominik A Klein and Simone Frintrop. 2011. Center-surround divergence of feature statistics for salient object detection. In *2011 International Conference on Computer Vision.* IEEE, 2214–2219.

[26] Baisheng Lai and Xiaojin Gong. 2016. Saliency guided dictionary learning for weakly-supervised image parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3630–3639.

[27] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV).* IEEE, 239–248.

[28] Guanbin Li and Yizhou Yu. 2015. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 5455–5463.

[29] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. 2014. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 280–287.

[30] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. 2019. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3917–3926.

[31] Nian Liu, Junwei Han, and Ming-Hsuan Yang. 2018. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3089–3098.

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3431–3440.

[33] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. 2016. Joint semantic segmentation and depth estimation with deep convolutional networks. In *2016 Fourth International Conference on 3D Vision (3DV).* IEEE, 611–619.

[34] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. 2012. Leveraging stereopsis for saliency analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 454–461.

[35] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. 2014. Rgbd salient object detection: a benchmark and algorithms. In *European conference on computer vision.* Springer, 92–109.

[36] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. 2019. Depth-Induced Multi-Scale Recurrent Attention Network for Saliency Detection. In *Proceedings of the IEEE International Conference on Computer Vision.* 7254–7263.

[37] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7479–7489.

[38] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. 2017. RGBD salient object detection via deep fusion. *IEEE Transactions on Image Processing* 26, 5 (2017), 2274–2285.

[39] Ling Shao and Michael Brady. 2006. Specific object retrieval based on salient regions. *Pattern Recognition* 39, 10 (2006), 1932–1948.

[40] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations.*

[41] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. 2019. Selectivity or invariance: Boundary-aware salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision.* 3799–3808.

[42] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. 2019. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 9799–9809.

[43] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. 2018. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2022–2030.

[44] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. 2017. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 136–145.

[45] Ningning Wang and Xiaojin Gong. 2019. Adaptive fusion for rgb-d salient object detection. *IEEE Access* 7 (2019), 55277–55284.

[46] Rui Wang, Stephen M Pizer, and Jan-Michael Frahm. 2019. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 5555–5564.

[47] Zhe Wu, Li Su, and Qingming Huang. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3907–3916.

[48] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. 2013. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1155–1162.

[49] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3166–3173.

[50] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. 2019. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE International Conference on Computer Vision*. 5684–5693.

[51] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. 2018. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1741–1750.

[52] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. 2018. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 714–722.

[53] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. 2019. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3927–3936.

[54] Ting Zhao and Xiangqian Wu. 2019. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3085–3094.

[55] Chunbiao Zhu and Ge Li. 2017. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 3008–3014.