Is Depth Really Necessary for Salient Object Detection?

Jiawei Zhao†, Yifan Zhao†, Jia Li*, Xiaowu Chen

†State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University

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 only takes the depth information as training data while only relies on RGB information 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. The code and model will be publicly available.

CCS CONCEPTS

• Computing methodologies → Interest point and salient region detections.

KEYWORDS

salient object detection, depth awareness, RGBD

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.

Combining 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 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 contrastive 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 constraint [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. [36] 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 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 convoluted 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 per-pixel vision tasks, such as semantic segmentation [33], surface normal [50].

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.

### 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 $f^a, f^b \in \mathbb{R}^{W' \times H' \times C'}$, we use a pixel-wise multiplication to enhance the common pixels in feature maps, while alleviates 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:

$$f^c = \xi_a(f^c) \otimes \xi_b(f^\beta) \otimes (\xi_a(f^c) \otimes \xi_b(f^\beta)),$$

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

After obtaining rich feature $f^c \in \mathbb{R}^{W \times H \times 3C'}$ 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 for the attention weights. The $f^c$ are then squeezed with a global average pooling, followed by a sigmoid normalization $\sigma$, and transformed as the vector shape to align the dimensions with feature channels. This serialized operation has the form:

$$g = \frac{1}{W' \times H'} \sum_{i=1}^{W'} \sum_{j=1}^{H'} f^c_{i,j},$$

$$u_{i,j} = f^c_{i,j} \otimes \sigma(\varphi(g_{i,j})).$$

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

$$G = \tau_d(\tau_c1(\xi_a(f^c)) \otimes \xi_a(u)) \otimes \tau_c2(\xi_b(f^\beta) \otimes \xi_b(u)),$$

where $\xi_{\{u1,u2\}}$ denotes the typical decoder and $\tau_{\{c1,c2,g\}}$ denotes the typical decoder with dimensional reductions as original input. Hence the relevant features to target object can be enhanced in the final output $G$. 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 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. This would 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 $e_{ij}$ of each pixel has the form:

$$e_{ij} = \frac{\sum_{h=1}^{h} \sum_{w=1}^{w} (\log p^d_{ij} - \log y^d_{ij})}{\sum_{h=1}^{h} \sum_{w=1}^{w} \max(\log p^d - \log y^d)}.$$

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.

![RGB images and depth maps](image)

**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 $p^s_i, y^s_i \in \mathbb{R}^{W \times H \times 1}$ be the predicted salient map and corresponding groundtruth, the SOD module is supervised with the BCE loss:

$$L_{bce} = -\sum_{i=1}^{H} \sum_{j=1}^{W} y_{ij}^s \log(p_{ij}^s).$$  \hspace{1cm} (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:

$$L_{iou} = 1 - \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} (y_{ij}^s \times p_{ij}^s) + 1}{\sum_{i=1}^{H} \sum_{j=1}^{W} (y_{ij}^s + p_{ij}^s) - y_{ij}^s \times p_{ij}^s + 1}. \hspace{1cm} (7)$$

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 $\epsilon$:

$$L_{depth} = \frac{1}{W \times H} \sum_{i=1}^{H} \sum_{j=1}^{W} || \log y_{ij}^d - \log p_{ij}^d ||^2_2. \hspace{1cm} (8)$$

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

$$L_{dec} = -\frac{\sum_{i=1}^{H} \sum_{j=1}^{W} \epsilon_{ij} \times y_{ij}^s \log(p_{ij}^s)}{\sum_{i=1}^{H} \sum_{j=1}^{W} \epsilon_{ij}}. \hspace{1cm} (9)$$

This error loss $L_{dec}$ 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:

$$L = L_{depth} + \sum_{i=1}^{3} \lambda_i (L_{bce} + L_{iou} + L_{dec}), \hspace{1cm} (10)$$

where $\lambda_i$ 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 $\lambda$ 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 ($F^\beta_{\text{max}}$), mean F-measure ($F^\beta_{\text{mean}}$), mean absolute error ($\text{MAE}$) and structure similarity measure ($S_o$) [15] on both RGB-based methods and RGBD-based methods.

We use $F^\beta$ to measure both Precision and Recall comprehensively. $F^\beta$ is computed based on Precision and Recall pairs as follows:

$$F^\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \hspace{1cm} (11)$$

where we set $\beta^2=0.3$ to emphasize more on Precision than Recall, and compute $F^\beta_{\text{max}}$, $F^\beta_{\text{mean}}$ using different thresholds as in [1].

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

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

where we set $\alpha=0.5$ 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 $\times$ 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.
Table 1: Performance comparison with 9 state-of-the-art RGBD-based SOD methods on five benchmarks. Smaller MAE, larger $F_{\text{max}}$, $F_{\text{mean}}$ and $S_a$ indicates better performance. The best results are highlighted in bold.

| methods | NUD-TE | NLPR-TE | STEREO | DES | SSD |
|---------|--------|---------|--------|-----|-----|
|         | $F_{\text{max}}$ | $F_{\text{mean}}$ | MAE | $S_a$ | $F_{\text{max}}$ | $F_{\text{mean}}$ | MAE | $S_a$ | $F_{\text{max}}$ | $F_{\text{mean}}$ | MAE | $S_a$ | $F_{\text{max}}$ | $F_{\text{mean}}$ | MAE | $S_a$ | $F_{\text{max}}$ | $F_{\text{mean}}$ | MAE | $S_a$ |
| DF [38] | .804 | .744 | .141 | .763 | .778 | .682 | .085 | .802 | .757 | .616 | .141 | .757 | .766 | .566 | .093 | .752 | .735 | .699 | .142 | .747 |
| AFNet [45] | .775 | .764 | .100 | .772 | .771 | .755 | .058 | .799 | .823 | .806 | .075 | .825 | .728 | .713 | .068 | .770 | .729 | .709 | .099 | .776 |
| CTMF [19] | .845 | .788 | .085 | .849 | .825 | .723 | .056 | .860 | .831 | .786 | .086 | .848 | .844 | .765 | .055 | .863 | .781 | .748 | .082 | .813 |
| 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 |
| CPF [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.

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. The PyTorch implementation will be publicly available.

4.3 Comparisons with the state-of-the-art

RGB-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], CPF [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 obviously that our method achieves a new performance leader-board with no depth image as inputs, which puts our model in inferior places for comparison. Especially for the $F_{\text{max}}$ and $F_{\text{mean}}$ 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 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].

---

1Link is masked for blind review policy.
Table 2: Performance comparison with 10 state-of-the-art RGB-based SOD methods on five benchmarks. Smaller $\text{MAE}$, larger $F_{\max}$, $F_{\text{mean}}$, and $S_\alpha$ correspond to better performance. The best results are highlighted in bold.

| methods       | ECSSD | DUT-TE | DUT-OMRON | HKU-IS | PASCAL-S |
|---------------|-------|--------|-----------|--------|----------|
|               | $F_{\max}$ | $F_{\text{mean}}$ | $\text{MAE}$ | $S_\alpha$ | $F_{\max}$ | $F_{\text{mean}}$ | $\text{MAE}$ | $S_\alpha$ | $F_{\max}$ | $F_{\text{mean}}$ | $\text{MAE}$ | $S_\alpha$ |
| BMPM [51]     | .929  | .894   | .045      | .911   | .911     | .851  | .762   | .049   | .801   | .774  | .698   | .064   | .808   | .921  | .875   | .039   | .905   | .862  | .803   | .073   | .840   |
| PAGR [52]     | .927  | .894   | .061      | .889   | .854     | .854  | .784   | .056   | .838   | .771  | .711   | .071   | .775   | .918  | .886   | .048   | .887   | .854  | .803   | .094   | .815   |
| R3Net [11]    | .929  | .883   | .051      | .910   | .829     | .816  | .716   | .067   | .837   | .793  | .690   | .067   | .819   | .910  | .853   | .047   | .894   | .837  | .775   | .101   | .809   |
| PiCA-R [31]   | .935  | .901   | .047      | .918   | .860     | .816  | .816   | .051   | .868   | .803  | .762   | .065   | .829   | .919  | .880   | .043   | .905   | .881  | .851   | .077   | .845   |
| BANet [41]    | .939  | .917   | .041      | .924   | .872     | .829  | .840   | .087   | .879   | .782  | .750   | .061   | .832   | .923  | .893   | .037   | .913   | .847  | .839   | .079   | .852   |
| PoolNet [30]  | .944  | .915   | .039      | .921   | .880     | .809  | .809   | .040   | .883   | .808  | .747   | .055   | .833   | .933  | .899   | .032   | .916   | .869  | .822   | .074   | .845   |
| BASNet [37]   | .943  | .880   | .037      | .916   | .859     | .791  | .791   | .048   | .866   | .805  | .756   | .056   | .836   | .928  | .895   | .032   | .909   | .857  | .775   | .078   | .832   |
| CPD-R [47]    | .939  | .917   | .037      | .918   | .865     | .805  | .805   | .043   | .869   | .797  | .747   | .056   | .825   | .925  | .891   | .034   | .905   | .864  | .824   | .072   | .842   |
| F3Net [24]    | .945  | .925   | .033      | .924   | .890     | .840  | .840   | .035   | .888   | .813  | .766   | .053   | .838   | .937  | .910   | .028   | .917   | .880  | .840   | .064   | .855   |
| GCPANet [7]   | .948  | .919   | .035      | .927   | .888     | .817  | .817   | .040   | .891   | .812  | .748   | .056   | .839   | .938  | .898   | .031   | .920   | .876  | .836   | .064   | .861   |
| Ours          | .950  | .932   | .032      | .927   | .896     | .853  | .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.

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.

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  | NIUD-TE $\text{MAE}$ |
|------|------|------|------|------|------|----------------------|
| ✔    |      | ✔    | ✔    | ✔    | ✔    | .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.

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. This more effective fusion strategy can improve $p^{\text{mean}}$ 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 $F_{\beta}$ 1.8% in total. Compared with the baseline using CAF and both BCE loss and IoU loss, our proposed DAM and DEC can improve $p^{\text{mean}}$ 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.

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.

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

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

| $\alpha$          | $\beta$ (max) | $\beta$ (mean) | MAE | $S_\alpha$ |
|-------------------|---------------|----------------|-----|------------|
| .924              | .929          | .925           | .929| .926       |
| .895              | .904          | .898           | .907| .904       |
| .024              | .021          | .022           | .021| .023       |
| .924              | .928          | .926           | .929| .926       |

Hyper-parameters. 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 $7 \times 7$ 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 constraints 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.

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, Lijiang Nie, Jain Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-ecn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE conference on...
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.