Learning Selective Mutual Attention and Contrast for RGB-D Saliency Detection

Nian Liu, Member, IEEE, Ni Zhang, Ling Shao, Senior Member, IEEE, and Junwei Han, Senior Member, IEEE

Abstract—How to effectively fuse cross-modal information is the key problem for RGB-D salient object detection. Early fusion and the result fusion schemes fuse RGB and depth information at the input and output stages, respectively, hence incur the problem of distribution gap or information loss. Many models use the feature fusion strategy but are limited by the low-order point-to-point fusion methods. In this paper, we propose a novel mutual attention model by fusing attention and contexts from different modalities. We use the non-local attention of one modality to propagate high-order and trilinear cross-modal interaction. We also propose to induce contrast inference from the mutual attention and obtain a unified model. Considering low-quality depth data may detriment the model performance, we further propose selective attention to reweight the added depth cues. We embed the proposed modules in a two-stream CNN for RGB-D SOD. Experimental results have demonstrated the effectiveness of our proposed model. Moreover, we also construct a new challenging large-scale RGB-D SOD dataset with high-quality, thus can both promote the training and evaluation of deep models.

Index Terms—salient object detection, RGB-D image, attention model, contrast, non-local network.

1 INTRODUCTION

O

VER the past few decades, researchers have proposed many computational salient object detection (SOD) models and achieved very promising performance, e.g., [1], [2], [3], [4], [5]. However, most of them work on RGB images and only leverage appearance cues, which usually incur insurmountable difficulties in many challenging scenarios. In the meantime, as human beings, we live in a real 3D environment. Our visual system heavily relies on depth information, which can supply sufficient complementary cues for the appearance information. Thus, it is quite natural and necessary to incorporate both RGB and depth data for solving the SOD problem.

To combine these two modalities for RGB-D SOD, some existing works employ the early fusion strategy [6], [7], [8], [9], with which SOD models take both RGB and depth data as inputs and process them in a unified way. However, this kind of model faces the difficulty of using one model to fit the data from two modalities well due to their distribution gap. Some other models [10], [11], [12] adopt two submodules to generate saliency maps for the RGB data and the depth data separately, and then use a fusion method to combine the two saliency maps. This kind of strategy is called result fusion. These methods are also suboptimal since using two separate SOD modeling processes will gradually compress and lose rich modality information. As a result, cross-modal interaction between the two saliency maps is highly limited.

Many other RGB-D SOD works exploit the middle fusion strategy as a better choice, which first fuses intermediate information of the two modalities and then generates a final saliency map. Most typically, many recent models [13], [14], [15], [16], [17], [18], [19], [20] first extract RGB and depth features separately using two-stream CNNs, and then combine the cross-modal features in decoders. We generalize these methods as belonging to the feature fusion strategy. Although it avoids the distribution discrepancy problem and fuses rich multi-modal features, the feature fusion methods are usually simple summation, multiplication, or concatenation, without exploring more powerful multi-modal feature interaction.

In this paper, we propose a novel and more effective middle fusion strategy. Inspired by the Non-local (NL) network [21], we propose to exploit cross-modal attention propagation. The NL network first generates long-range spatial attention for each query position via computing the query-key pair-wise affinity, and then uses the attention to propagate global context features. Given that the attention and the propagated features are both induced from the same feature map, the NL model is regarded as belonging to the self-attention mechanism. However, since the attention mainly activates on regions similar to the query position, the NL model is very sensitive to the feature quality. If the input feature map is not discriminative enough, the propagated context features from similar regions can only supply limited information gain for each query position. An intuitive example is given in Figure 1. We can see from (c) that the RGB feature map only has large activations on the vase and ignores most parts of the flower. Hence, for a query position located on the body of the vase, which is marked as the white point in (e), its self-attention map mainly activates on the vase. As a result, the feature map after using self-attention shown in (g) and the resultant saliency map shown in (i) still miss most parts of the flower.

Considering the property that RGB and depth data can complement each other, we propose to propagate global context using each other’s attention, to which we refer as the mutual-attention mechanism. It supplies complementary cues about where should
We construct a large-scale RGB-D SOD dataset with diverse real-life scenes and good depth quality. We have released the dataset\(^1\) and believe it can benefit the RGB-D SOD research community much.

To sum up, the contributions of our work can be summarized as follows:

- We present a novel mutual attention model for multi-modal information fusion in RGB-D SOD. It leverages complementary attention knowledge in long-range context propagation and introduces high-order and trilinear modality interactions. We also propose to unify the contrast mechanism in it.
- We adopt the proposed model in a two-stream UNet for RGB-D SOD. In the decoders, mutual attention can be further used for multi-scale cross-modal interaction. We also propose selective attention to weight the fused depth cues thus reduce the distraction of low-quality depth information.
- Aiming at the problem of lack of high-quality and large-scale benchmark datasets, we construct a new RGB-D SOD dataset with the most image-pairs, the most diverse visual scenes, and high-quality depth maps. It can help train deep models with better generalization and achieve a more comprehensive evaluation.
- We conduct performance evaluation on nine RGB-D SOD benchmark datasets. Experimental results verify the effectiveness of our proposed models, especially the mutual attention model. Finally, our overall RGB-D SOD model performs favorably against previous state-of-the-art methods.

Compared with our previous version of this work\,[32], we have made the following extensions. First, we found self-attention was not beneficial when fused with mutual attention as in\,[32], which we argue is due to the low-quality input feature maps. Thus we do not fuse self-attention and mutual attention anymore. Instead, we fuse mutual attention and contrast. Then we find it useful to cascade a self-attention model right after using mutual attention and contrast. We believe it is because the feature maps have been largely promoted and become more discriminative after using mutual attention and contrast. Second, in\,[32] we only used the S\(^2\)MA module right after the encoders. In this work, we find that further using mutual attention in subsequent decoders can improve the model performance. Third, we propose a new and more effective selective attention method in this work and adopt it in every cross-modal fusion module. Forth, we construct a new large-scale, and high-quality RGB-D SOD benchmark dataset.

\(^1\)https://github.com/nnizhang/SMAC

## 2 Related Work

### 2.1 Saliency Detection on RGB-D Images

Traditional RGB-D SOD methods usually borrow common priors (e.g., contrast\,[28] and compactness\,[33]) from RGB SOD models to design RGB and depth features. Additionally, some researchers proposed to exploit depth-specific priors, e.g., shape and 3D layout priors\,[34], and anisotropic center-surround difference\,[11], [30].

The aforementioned methods all rely heavily on hand-crafted features and lack high-level representations, which are very important for understanding challenging scenarios. To this end, many recent works introduce CNNs into RGB-D SOD and have achieved promising results. Qu et al.\,[7] adopted the early fusion strategy and serialized hand-crafted RGB and depth features together as the CNN inputs. Fan et al.\,[9] and Liu et al.\,[8] used each depth map as the \(4^{th}\) channel along with the corresponding RGB channel along with the corresponding RGB.
image as the CNN input. In [12], Wang et al. adopted the result fusion strategy and adaptively fused RGB and depth saliency maps with a learned switch map. Recently, the middle fusion strategy is adopted by many works to fuse intermediate depth and appearance features. Han et al. [16] fused the fully-connected representations of the RGB and the depth branches into a joint representation. Most other models [13], [17], [18], [19], [35] fused complementary cross-modal convolutional features at multiple scales by various methods, such as summation, multiplication, and concatenation. In contrast, our model fuses cross-modal non-local attention and context features, thus introducing high-order and trilinear information interactions.

The attention mechanism is also widely used in existing works to fuse RGB and depth modalities, e.g., in [36] and [13]. However, they only generated channel [13] or spatial [36] attention from the depth view and adopted them to filter the appearance features. Nevertheless, we generate non-local attention from both views and then use them to propagate long-range contexts for each other.

### 2.2 Self-Attention and Multi-modal Attention

In [37], Vaswani et al. proposed a self-attention network for natural language modeling. Given a query word and a set of key-value pairs for other words, they computed attention weights for all query-key pairs and then aggregated all the values as the context feature. Similarly, Wang et al. [21] proposed the NL model for learning self-attention in 2D or 3D vision modeling. For multi-modal attention learning, Nam et al. [38] proposed to learn both visual and textual attention mechanisms for multi-modal reasoning and matching. Wan et al. [39] applied three attention models in three modalities of source code for the code retrieval task. However, both of them learn and adopt attention for each modality separately, and then fuse the obtained attended features. In [40] and our previous work [32], cross-task and cross-modal attention affinities are fused. Different from them, we propose the mutual attention mechanism to mix up attention and values from different modalities. We also find it more helpful to cascade self-attention right after mutual attention instead of fusing their affinities.

### 2.3 RGB-D SOD Datasets

So far, researchers have proposed several RGB-D saliency datasets and eight of them are widely used in RGB-D SOD papers. Niu et al. [26] collected and annotated the first stereoscopic SOD dataset STERE. It contains 1,000 pairs of stereoscopic Internet images, most of which have various outdoor scenes and objects. The depth maps are generated using the SIFT flow [41] algorithm. Li et al. [27] constructed a light field SOD dataset LFSD using the Lytro light field camera. This dataset includes 100 indoor and outdoor images and the depth maps are directly generated by the Lytro desktop. Most images contain simple foreground objects but have complex or similar backgrounds. The NLPR [29] dataset and the RGBDI35 [28] dataset contain 1,000 and 135 images collected by the Microsoft Kinect, respectively. Hence, their depth maps have good quality. However, most of their images include relatively simple foreground objects and visual scenes. The NJUD [30] dataset has 1,985 stereo images collected from the Internet, 3D movies, and photos taken by a Fuji W3 camera. Most images show diverse outdoor scenes and foreground objects. The depth maps are generated using Sun’s optical flow method [42]. SSD [31] is a small-scale dataset which only has 80 stereo movie frames. The images contain several movie scenes with persons, animals, buildings, etc., as foreground objects. The depth maps are also generated using [42]. DUTLF-Depth [13] is a recently proposed dataset that contains 1,200 images captured by a Lytro2 light field camera. It includes diverse complex scenes, e.g., multiple or transparent objects, complex backgrounds, and low-intensity scenes. The authors adopted the method of [43] to obtain the depth maps. Fan et al. [9] proposed a Salient Person (SIP) dataset with 929 images to emphasize persons in real-world scenes. All the images and depth maps are collected by a Huawei Mate10 smartphone using its dual camera. A summarization of these datasets is shown in Table 1.

Although eight benchmark datasets seem many enough for the RGB-D SOD research, we argue that most of them are unsatisfactory due to three points. First, they do not have diverse enough visual scenes for effectively training and comprehensively evaluating SOD models. Many of them only have simple salient objects and similar background scenes. Second, most of them have insufficient images, thus being unsuitable for training modern deep learning-based models. From Table 1 we can see that only two in the eight datasets have more than 1,000 images. Hence most works use two or three datasets together to train deep networks. Third, since many stereo image-based datasets used early flow-based algorithms to generate depth maps, their depth map quality is usually low due to inaccurate pixel matching, noises in faraway areas, and over-smooth segmentation. On the contrary, the proposed dataset totally includes 3,179 images with various visual scenes and high-quality depth maps, which can promote both training and evaluation of deep RGB-D SOD models.

### 3 Proposed Dataset

In this paper, we construct a new large-scale challenging RGB-D SOD dataset based on the ReDWeb [47] dataset. It is a state-of-the-art dataset proposed for monocular image depth estimation. It contains 3,600 images selected from many web stereo images with various real-life scenes and foreground objects. For depth map generation, the authors first adopted the state-of-the-art FlowNet2.0 algorithm [46] to generate correspondence maps, and then used a deep semantic segmentation model [48] to remove noises in sky areas as the post-process, thus resulting in high-quality depth maps. Consequently, ReDWeb supplies a good basis for constructing our high-quality RGB-D SOD dataset, which we name as ReDWeb-S. We elaborate on the dataset construction method and the statistic analysis in this part.

#### 3.1 Dataset Construction

We first ask four participants to annotate salient objects in each image of the ReDWeb dataset using bounding-boxes (bbox) and also remove images without foreground objects. Then, for each image, we calculate the IoU matching scores for the bboxes of every two annotators and select the bbox with the highest total matching score as the annotation. At last, we ask participants to accurately segment salient objects based on the selected bboxes. As a result, we obtain 3,179 images with both high-quality depth maps and annotated saliency maps. We further randomly split them into a training set with 2,179 RGB-D image pairs and a testing set with the remaining 1,000 image pairs.

### 3.2 Dataset Statistics and Comparisons

In this part, we analyze the proposed ReDWeb-S dataset from several statistical aspects and also conduct a comparison between
**Table 1**

| Datasets   | Year | Scene Stat | Depth Stat | GC | IC |
|------------|------|------------|------------|----|----|
|            |      | IN | SN | ON | Sensor | DQ | RGC | DGC | RIC | DIC | CBI | OS |
| STERE [26] | 2012 | 1000 | 243 | 310 | Stereo images+sift flow [41] | 0.9770 | 287.35 | 407.16 | 4.22 | 1.13 | 103.76 | 0.21 |
| LFSD [27]  | 2014 | 100 | 52 | 63 | Lytro light field camera [44] | 0.9826 | 230.10 | 290.36 | 4.37 | 1.01 | 118.30 | 0.27 |
| RGBD135 [28] | 2014 | 135 | 48 | 41 | Microsoft Kinect [45] | 0.9866 | 51.12 | 136.39 | 3.61 | 0.89 | 80.86 | 0.13 |
| NLPR [29]  | 2014 | 1000 | 175 | 237 | Microsoft Kinect [45] | 0.9896 | 148.78 | 42.59 | 3.54 | 1.01 | 84.16 | 0.13 |
| NJUD [30]  | 2014 | 1985 | 282 | 359 | Stereo image+Sun’s [42] | 0.9852 | 152.31 | 254.55 | 4.22 | 1.29 | 114.32 | 0.24 |
| SSD [31]   | 2017 | 80 | 38 | 23 | Stereo movies+Sun’s [42] | 0.9845 | 494.92 | 796.82 | 3.42 | 1.20 | 139.44 | 0.21 |
| DUTLF-Depth [13] | 2019 | 1200 | 191 | 291 | Lytro2 light field camera+ [43] | 0.9852 | 18.53 | 44.48 | 3.94 | 1.29 | 123.29 | 0.23 |
| SIP [9]    | 2019 | 929 | 69 | 1 | Huawei Mate10’s dual camera | 0.9920 | 98.03 | 79.90 | 3.05 | 0.73 | 106.25 | 0.20 |
| ReDWeb-S (Ours) | - | 3179 | 332 | 432 | Web stereo images+Flownet2.0 [46]+post-processing | 0.9904 | 82.78 | 53.10 | 4.38 | 1.60 | 148.72 | 0.27 |

ReDWeb-S and other existing RGB-D SOD datasets, as shown in Table 1.

**Image Numbers**: This is a very important factor for modern data-hungry deep learning-based SOD models. From Table 1, we can see that previous datasets have at most no more than 2,000 images and most of them have less than 1,000 images. Our ReDWeb-S dataset has a more than 1,000 increase and becomes the largest RGB-D SOD benchmark dataset.

**Scene Statistics**: Rich scene and object categories are beneficial for both promoting and evaluating the generalization ability of SOD models. To evaluate the scene richness, we use a state-of-the-art ResNet-50 [49] based scene classification model pretrained on the Places365 [50] dataset to conduct scene categorization for each image in each dataset. Finally, we count the total scene class number for each dataset and report them in Table 1. We can observe that our ReDWeb-S dataset has the most scene class number among all the nine datasets and it is much larger than those of most of the others. Similarly, we also evaluate the object richness of these datasets by conducting object classification. Specifically, we adopt a state-of-the-art ResNeXt-101 [51] based image classification model [52] to classify each image into one of the 1000-class object labels defined in ImageNet [53]. One thing to notice is that ImageNet labels do not include the “person” class, which is very commonly seen in RGB-D SOD datasets. Therefore, we first manually pick out images whose foreground objects are persons and perform object classification on the remaining images. At last, we report the total object class number of each dataset in Table 1. The results show that our dataset has the most diversiform objects. To have a deeper look at the scene and object diversities of our ReDWeb-S dataset, we also show its distributions of the top 60% scene and object categories in Figure 2. We can see that different scene categories have an approximately smooth distribution. However, for the object class distribution, we observe that nearly 20% images of our ReDWeb-S dataset belong to the “person” category, which is a dominant proportion in the distribution histogram. This is reasonable since “person” is often the leading role in real-life photos and we also observe similar phenomena in other RGB-D SOD datasets, such as NJUD [30], SSD [31], and STERE [26].

**Depth Map Quality**: Since depth maps provide very important complementary cues for saliency detection, their quality is also of significant importance. Depth maps with higher quality can supply more accurate guidance information for RGB-D SOD. Hence, we evaluate the depth map quality (DQ) based on the bad point rate (BPR) proposed in [54]. BPR is a state-of-the-art no-reference depth assessment metric and is calculated as the proportion of mismatched pixels between depth edges and texture edges. We define the DQ score as $1 - BPR$, which is the matching
Fig. 3. Comparison of nine RGB-D SOD dataset in terms of the distributions of global contrast and interior contrast.

Fig. 4. Comparison of the average annotation maps for nine RGB-D SOD benchmark datasets. We also use a 2D Gaussian distribution to fit each map and mark the corresponding coordinates of the center point ($\mu_x$ and $\mu_y$) and the standard deviations ($\sigma_x$ and $\sigma_y$).

accuracy. From Table 1 we can see that SIP has the best depth quality while ReDWeb-S ranks the second. We also observe that early datasets such as STERE and LFSD have the worst depth quality since they use the oldest depth estimation methods.

**Global Contrast:** Since global contrast can help localize salient objects, it can be used to assess the challenging of each dataset. We first evaluate the global contrast for the RGB modality, denoted as “RGC” in Table 1. We follow [55] to compute the $\chi^2$ distance between the RGB color histograms of foreground and background regions for each image as the measurement of RGC. Finally, we report its average value for each dataset in Table 1. Similarly, we also report the measurements of the global contrast for the depth modality (denoted as “DGC”) by computing the $\chi^2$ distance between depth histograms. The results demonstrate that ReDWeb-S has a relatively small global contrast. We also show the detailed distributions of RGC and DGC in Figure 3. Since the data ranges of the computed RGC and DGC are very large, we coarsely divide them into four scopes, i.e., very low, low, medial, and high. We observe that our proposed ReDWeb-S dataset mainly have low and medial RGC, and very low, low, medial DGC. These results clearly demonstrate its challenging for RGB-D SOD.

**Interior Contrast:** We refer to “interior contrast” as the contrast within the salient object of each image. A small value means the different parts of a salient object have homogenous appearance, thus making it easier to uniformly highlight the whole object. On the contrary, a large value indicates the salient object is more sophisticated and harder to detect. We evaluate the interior contrast by computing the information entropy of the normalized color and depth histograms of foreground regions for the RGB and depth modality, respectively. The average RGB interior contrast (RIC) and depth interior contrast (DIC) of each dataset are reported in Table 1. We find that ReDWeb-S achieves the largest average RIC and DIC. We also draw the curves of the distributions of RIC and DIC for all the datasets. Figure 3(c) shows that ReDWeb-S and NJUD both have the most images with large RIC. Figure 3(d) indicates ReDWeb-S has much more images with high DIC than other datasets. These observations further verify the difficulty of our proposed ReDWeb-S dataset.

**Center Bias:** Center bias is a well-known prior knowledge for saliency detection since people usually put target objects in the middle of their photos. We follow previous works (e.g., [56]) to draw the average annotation map (AAM), i.e., the average of ground-truth annotations of all images, on each dataset to illustrate their center biases in Figure 4. We resize each AAM to $256 \times 256$ pixels to ease the comparison. For quantitatively analyzing the degree of center bias of each dataset, we propose to normalize each AAM and then use a 2D Gaussian distribution to fit it since most of them clearly show similar patterns with 2D Gaussian surfaces. Then, we mark the coordinates of the center point ($\mu_x$ and $\mu_y$) and the standard deviations ($\sigma_x$ and $\sigma_y$) along the width and the height direction on each AAM in Figure 4. For $\mu_x$ and $\mu_y$, we report their offsets to the map center to ease understanding. Usually, larger center point offsets and standard deviations mean that one AAM deviates its center more, thus this dataset has less center bias. Among them, standard deviations are more important since the center points of most AAMs are close to the map centers and large standard deviations indicate salient objects spread in a large range on this dataset. From Figure 4,
we observe that ReDWeb-S has moderate center point offsets and the largest standard deviations, demonstrating it shows less center bias. We also compute the average of $\sigma_x$ and $\sigma_y$ as a center bias index (CBI) to comprehensively assess the degree of center bias for each dataset and report them in Table 1. The comparison shows that ReDWeb-S achieves the largest CBI.

Object Size: The sizes of salient objects also play an important role in SOD since usually both very large or small salient objects are difficult to detect. We compute the normalized object size for each image and draw its distribution for each dataset in Figure 6. It shows that most previous datasets usually have more small objects while ReDWeb-S has more large ones. The former is difficult to locate while it is hard to completely highlight the latter. We also report the average OS of each dataset in Table 1. The results show that ReDWeb-S and LFSD achieve the largest average object size.

Some example images of the ReDWeb-S dataset are given in Figure 5. We can see that the depth maps are of high quality and there are various challenging scenarios, such as complex foreground and backgrounds, low-contrast images, and transparent objects.

Fig. 5. Example images of our proposed ReDWeb-S dataset.

Fig. 6. Comparison of the distribution of object size (OS) for nine datasets.

In this section, we elaborate on the proposed Selective Mutual Attention and Contrast (SMAC) module for fusing multi-modal information in RGB-D SOD. It is built based on the NL module [21] and the whole network architecture is shown in Figure 7. We first briefly review the NL module and then go into our SMAC module.

4.1 Reviewing the NL module

Considering a feature map $X \in \mathbb{R}^{H \times W \times C}$, where $H$, $W$, and $C$ represent its height, width, and channel number, respectively, the NL module first embeds $X$ into three feature spaces with $C'$ channels:

$$\theta(X) = X W_{\theta}, \phi(X) = X W_{\phi}, g(X) = X W_{g},$$  \hspace{1cm} (1)

where $W_{\theta}, W_{\phi},$ and $W_{g} \in \mathbb{R}^{C \times C'}$ are the embedding weights in the query, key, and value spaces, respectively. They can be implemented using $1 \times 1$ convolutional (Conv) layers.

Next, a similarity (or affinity) function $f$ is adopted using $\theta$ and $\phi$ embeddings as inputs. In [21], the authors have proposed several forms for the function $f$. Here we adopt the most widely used dot product function, i.e.,

$$f(X) = \theta(X) \phi(X)^T,$$  \hspace{1cm} (2)

where $f(X) \in \mathbb{R}^{HW \times HW}$. In $f(X)$, each element $f_{i,j}$ represents the affinity between the query position $i$ and the key position $j$ in $X$.

Subsequently, the NL module adopts normalization along the second dimension to generate an attention weight matrix:

$$A(X) = \text{softmax}(f(X)),$$  \hspace{1cm} (3)

where each row $A_i$ indicates the normalized attention of all key positions respect to the $i^{th}$ query position. Then, the values in $g$ are aggregated by weighted sum:

$$Y = A(X) g(X),$$  \hspace{1cm} (4)

where $Y \in \mathbb{R}^{HW \times C'}$ is an attentive feature. By decomposing the computation for each query position, we have:

$$Y_i = \sum_{j=1}^{HW} A_{i,j} \cdot g_j.$$  \hspace{1cm} (5)

We can see that for each position in $Y$, the obtained attentive feature is an aggregation of features at all positions. Thus $Y$ incorporates long-range relations within the whole feature map $X$.  

4 PROPOSED SELECTIVE MUTUAL ATTENTION AND CONTRAST MODEL

In this section, we elaborate on the proposed Selective Mutual Attention and Contrast (SMAC) module for fusing multi-modal information in RGB-D SOD. It is built based on the NL module [21] and the whole network architecture is shown in Figure 7. We
query, key spaces and obtain their attention matrixes: $R \times X \times RGB-D SOD$. Imaging we have two feature maps of different modalities to introduce context complementarity. achieved (see the experimental results in Section 6.3). For multi-result, very limited feature promotion and performance gain can be with self-similarity but without extra informative contexts. As a low-quality, the non-local attention can only attend to regions the quality of the original feature. If the original feature is of itself. Thus, it belongs to the self-attention category. We argue that the effectiveness of such a transform is highly limited by it. Therefore, we regard the depth modality as complementary information and have two feature maps coming from the RGB and the depth modality, respectively. $C \in \mathbb{R}^{C \times C}$ is the weight of a convolutional layer for projecting the attentive feature back to the original feature space.

**4.2 Mutual Attention**

The obtaining of the attentive feature $Y$ in the NL module can be rewritten as:

$$Y = \text{softmax}(XW_\theta W_\tau^\top X^\top)XW_g.$$  

(7)

We can see that it is a trilinear transform of the original feature $X$ itself. Thus, it belongs to the self-attention category. We argue that the effectiveness of such a transform is highly limited by the quality of the original feature. If the original feature is of low-quality, the non-local attention can only attend to regions with self-similarity but without extra informative contexts. As a result, very limited feature promotion and performance gain can be achieved (see the experimental results in Section 6.3). For multimodal tasks, such as RGB-D SOD, we can leverage the attention of different modalities to introduce context complementarity.

In this paper, we propose using Mutual Attention (MA) for RGB-D SOD. Imaging we have two feature maps $X^r, X^d \in \mathbb{R}^{H \times W \times C}$ coming from the RGB and the depth modality, respectively, we first follow the NL module to embed them into the query, key spaces and obtain their attention matrices:

$$A^r(X^r) = \text{softmax}(\theta^r(X^r)\phi^r(X^r)^\top),$$

$$A^d(X^d) = \text{softmax}(\theta^d(X^d)\phi^d(X^d)^\top).$$  

(8)

Then, we fuse multi-modal information by mixing the value modality and the attention modality up:

$$Y^r = A^d(X^d)g^r(X^r),$$

$$Y^d = A^r(X^r)g^d(X^d).$$  

Here the two modalities provide attention for each other, thus we refer to this attention scheme as mutual attention. It propagates cross-modal long-range contextual dependencies, which is a novel way for fusing multi-modal information.

By omitting the embedding weights and the Softmax normalization, we can rewrite the position-wise attentive features for the mutual attention as:

$$Y^r_i = \sum_{j=1}^{HW} <X^r_i, X^d_j>_r,$$

$$Y^d_i = \sum_{j=1}^{HW} <X^r_j, X^d_i>_d,$$  

(10)

where $<,>$ denotes the inner product of two feature vectors, and $X^r_i$ means the feature vector at position $i$. Reviewing previous widely used modality fusion methods such as summation, multiplication, and concatenation, they only involve point-to-point low-order fusion. We observe from (10) that our mutual attention module introduces high-order and trilinear interactions between $X^r$ and $X^d$, thus being able to explore more complex cross-modal information interaction.

**4.3 Incorporating the Contrast Mechanism**

The above attention mines spatial affinity via the feature inner product, thus integrating contexts with similar features from both views. Considering the effectiveness of the widely used contrast mechanism in SOD, which devotes to find the difference between salient regions and backgrounds, in this paper we also propose to incorporate the contrast mechanism. Reviewing that (2) computes the spatial affinity, we can easily obtain spatial dissimilarity by taking its opposite and then calculate a contrast attention:

$$C(X) = \text{softmax}(\frac{-f(X)}{T}),$$  

(11)

where the temperature $T$ is a learnable parameter to help learn an adaptive distribution for the contrast attention. Then, we can use this attention to aggregate features from contrastive regions and compute the feature difference as contrast. When integrating it with the proposed mutual attention, we can obtain a unified mutual attention and contrast (MAC) model and compute the modality-specific outputs $Z^r$ and $Z^d$ as:

$$Z^r = X^r + A^d(X^d)g^r(X^r)W^r_z - C^d(X^d)g^r(X^r)W^r_c,$$

$$Z^d = X^d + A^r(X^r)g^d(X^d)W^d_z - C^r(X^r)g^d(X^d)W^d_c,$$  

(12)

(13)

where $W^*_z \in \mathbb{R}^{C \times C}$ plays a similar role with $W^*_z$.

**4.4 Selective Attention**

The MAC model treats the information from the two modalities equally. However, considering that SOD using RGB data only has achieved very promising results yet [4], [5], [57] and several RGB-D SOD benchmark datasets have low-quality depth maps, we regard the depth modality as complementary information and
propose to reweight the depth cues using adaptive selective attention weights. In our previous work [32], we proposed the pixel-wise selective attention for both modalities. However, the attention weights are inferred by simply concatenating both RGB and depth features. Such an implicit and pixel-wise learning method makes the training no easy and we found it hard to work for our new MAC model. In this work, we propose a novel method to explicitly infer image-wise selective attention, which is more effective for our new model and the attention weight can be used for multiple cross-modal fusion modules. Inspired by recent monocular depth estimation models, we argue that depth information can be roughly estimated by the RGB features of each image. If the estimation error is large, it suggests the original depth map is probably of low-quality. Thus, we can infer the selective attention from the estimation error and use it to weight the incorporated depth cues.

Specifically, we first deploy two $1 \times 1$ Conv layers with $C$ channels on the top of $X^r$ to estimate $X^d$. Then, the estimation error map can be computed by subtracting the estimated $X^d$ from the real one. Next, two $1 \times 1$ Conv layers with setting stride = 2 are used to downsample the error map. We also reduce the channel numbers by setting them to 256 and 128, respectively. Batch normalization [58] and the ReLU activation function are used right after the first and the last two Conv layers, as shown in Figure 7. Finally, two FC layers with 256 and 1 nodes are used to predict the final selective attention $\alpha$, with adopting the Sigmoid activation function. The whole process can be roughly represented as:

$$\alpha = Sigmoid(FC(Conv(X^d - Conv(X^r)))) .$$  \hspace{1cm} (14)

Then, we can use the selective attention to weight the mutual attention and the contrast terms in (12) since they are induced from the depth modality and may suffer from the low-quality of depth maps:

$$Z^r = X^r + \alpha \cdot (A^d(X^d)g^r(X^r)W^r_z - C^d(X^d)g^r(X^r)W^r_f) .$$  \hspace{1cm} (15)

5 RGB-D SOD NETWORK

Based on the proposed SMAC model, we propose a novel deep model for RGB-D SOD. As shown in Figure 8(a), our model is based on a two-stream CNN, and each of them is responsible for generating the saliency map from the input modality based on the U-Net [25] architecture.

Specifically, we share the same network structure for the two encoders and adopt the VGG-16 network [59] as the backbone. We follow [60] to slightly modify its network structure as follows. First, we change the pooling strides of the pool4 and pool5 layers to 1 and set the dilation rates [61] of the conv5 block to 2. Second, we transform the fc6 layer to a $3 \times 3$ Conv layer with 1024 channels and set its dilation rate to 12. Similarly, we turn the fc7 layer into a $1 \times 1$ Conv layer with 1024 channels. As such, the encoder network becomes a fully convolutional network [62] with the output stride of 8, thus preserving large spatial resolutions for high-level feature maps.

Next, we adopt DenseASPP [63] modules on top of the two encoders for further enhancing their capability. DenseASPP introduces dense connections [64] to the ASPP [61] module and therefore covers dense feature scales. Before adopting DenseASPP, we first use $1 \times 1$ Conv layers to compress the two fc7

---

Fig. 8. Architecture of our proposed RGB-D SOD network. (a) shows the main two-stream network. The skip-connected VGG layers are marked in the first stream by “C”, “*” and “fc*”. The channel numbers of the feature maps are also marked in the second stream. “NL” means the Non-local estimation error and use it to weight the incorporated depth cues.
feature maps to 512 channels. Considering the specific training image size of our SOD network, we design three dilated Conv branches with dilation rates of 2, 4, and 8, respectively. All of them use $3 \times 3$ Conv kernels and 176 channels. Following [63], dense connections are deployed within the three branches. To capture the global context, we additionally design a fourth branch that average pools the input feature map first and then upsamples the result to the original size. At last, we concatenate the original input feature map and the outputs of the four branches, and then compress them to 512 channels. The whole module architecture is shown in Figure 8(b).

After the DenseASPP module, we take the output features of the RGB and depth streams as inputs and adopt the proposed SMAC module to perform cross-modal information interaction. Since the outputs $Z_r$ and $Z_d$ have leveraged context propagation and contrast inference from the cross-modal attention, their quality and discriminability have been improved. Hence, we further use an NL module for each of them to blend the received cross-modal cues, as shown in Figure 8(a).

Next, we go into the decoder parts. We represent the following decoder modules of the two branches as $\mathcal{D}$ and $\mathcal{D}'$, respectively. As shown in Figure 8(c), for each decoder module, we first follow the UNet [25] architecture to progressively fuse an intermediate encoder feature map with the previous decoder feature map. The used intermediate VGG features are the last Conv feature maps of the five blocks, which are marked in Figure 8(a). For encoder-decoder feature fusion, we simply concatenate them together and then adopt two Conv layers. To enhance cross-modal information interactions, we also deploy cross-modal fusion structures between $\mathcal{D}$ and $\mathcal{D}'$. Concretely, for the first three of them, we use the proposed selective mutual attention (SMA) modules. Here we do not incorporate contrast anymore to save computational costs. For the last two decoder modules, we choose to not adopt the SMA modules since they are computationally prohibitive for large feature maps. As an alternative, we simply concatenate cross-modal decoder features and then use another Conv layer to learn a residual fusion signal for the RGB branch. Note that here the fusion is unidirectional since we consider RGB information as the main character. The selective attention is also adopted for this residual fusion.

Each Conv layer in our decoder part uses $3 \times 3$ kernels and is followed by a BN [58] layer and the ReLU activation function. We set the output channel number of each decoder module to be the same as that of the next skip-connected VGG feature map, as marked in Figure 8(a). For the last three decoder modules, we upsample previous decoder feature maps since they have smaller spatial sizes than the skip-connected VGG feature maps. Hence, the decoder feature maps are progressively enlarged. Due to the third $\mathcal{D}$ and $\mathcal{D}'$ have relatively large spatial sizes, i.e., $\frac{1}{4}$ of the input size, we downsample the $\phi$ and $g$ feature maps in SMA by a factor of 2 using max-pooling to save computational costs. For saliency prediction, we adopt a $3 \times 3$ Conv layer with 1 channel on the last decoder feature map and use the Sigmoid activation function to obtain the saliency map for each CNN stream.

6 EXPERIMENTS

6.1 Evaluation Metrics

Following recent works, we adopt four metrics for a comprehensive evaluation. The first one is the F-measure score, which treats SOD as a binary classification task and comprehensively considers both precision and recall. We follow most previous works to report the maximum F-measure (maxF) score under the optimal threshold. The second metric is the Structure-measure $S_m$ [65]. It considers structural similarities between the saliency maps and the ground truth from the perspectives of both region-level and object-level. The third metric is the Enhanced-alignment measure $E_{E}$ [66] which jointly evaluates both global statistics and local pixel matching. The last metric we use is the widely used Mean Absolute Error (MAE). It is computed as the average of the absolute difference between a saliency map and the corresponding ground truth.

6.2 Implementation Details

Recent works [13], [19], [32], [67] train deep RGB-D SOD models using images from three datasets, i.e., 1,400 images of the NJUD dataset, 650 images of the NLPR dataset, and 800 images of the DUTLF-Depth dataset. We follow this criterion to train our network. We set the training and testing image size of our RGB-D SOD network as $256 \times 256$. Random cropping and horizontal flipping are adopted for data augmentation. As for the former, we resize each training image and the corresponding depth map to $288 \times 288$ pixels and then randomly crop $256 \times 256$ image regions as the network inputs. For the depth stream of our network, we replicate each single-channel depth map thrice to fit its input layer. Considering that different datasets have different depth presentations, we preprocess their depth maps to a unified presentation, i.e., small depth values indicate the pixels are close to the camera and vice versa. The depth maps are also normalized to the value range of $[0,255]$ to ease the network training. Before feeding into the two-stream network, each image and the corresponding three-channel depth map are subtracted by their mean pixel values as preprocessing.

We adopt the cross-entropy loss to train both streams of our SMAC network. Following previous works, deep supervision is also adopted in each decoder module to facilitate the network training. By following [60], we empirically use 0.5, 0.5, 0.8, 0.8, and 1 as the loss weights of the five decoder modules of each stream. We train our SMAC network totally using 40,000 iterations with the stochastic gradient descent (SGD) with momentum algorithm, where the initial learning rate, weight decay, momentum, and batchsize are set to 0.01, 0.0005, 0.9, and 12, respectively. We decay the learning rate by dividing it by 10 at the 20,000th and the 30,000th training steps, respectively.

Our SMAC model is implemented using the Pytorch [68] package. A GTX 1080 Ti GPU is used for computing acceleration. When testing, we resize each image pair to $256 \times 256$ pixels as the network inputs and use the outputted saliency map from the RGB stream as the final saliency map, without using any post-processing technique. Our SMAC network only takes 0.059 seconds for testing each image-pair.

6.3 Component Analysis

For a more comprehensive evaluation, we additionally add our ReDWeb-S dataset for training in the component analysis experiments. However, due to space limitation and the ease of comparison, we only report the comparison of different model settings on three challenging datasets, i.e., ReDWeb-S, NJUD, and LFSD. The experimental results are shown in Table 2.

Effectiveness of Different Model Components. The first row in Table 2 denotes the baseline model, i.e., adopting UNet and
Component analysis on the effectiveness of the proposed SMAC RGB-D SOD model. We first show the comparison among different model settings of gradually using the proposed model components. Blue indicates the best performance among these settings. We also compare our model with some other methods to verify its effectiveness, as shown in rows VII to XII.

### TABLE 2

| ID | Settings | RedWeb-S | NJUD [30] | LFSD [27] |
|----|----------|----------|-----------|-----------|
|    |          | $S_m$ | maxF | $E_\bar{y}$ | MAE | $S_m$ | maxF | $E_\bar{y}$ | MAE | $S_m$ | maxF | $E_\bar{y}$ | MAE |
| I  | Baseline | 0.772  | 0.760 | 0.833 | 0.109 | 0.883 | 0.873 | 0.921 | 0.055 | 0.834 | 0.819 | 0.877 | 0.088 |
| II | +MA      | 0.784  | 0.769 | 0.838 | 0.115 | 0.898 | 0.895 | 0.934 | 0.054 | 0.862 | 0.864 | 0.901 | 0.077 |
| III| +MAC     | 0.787  | 0.772 | 0.842 | 0.113 | 0.901 | 0.896 | 0.936 | 0.052 | 0.862 | 0.863 | 0.902 | 0.078 |
| IV | +MAC+NL  | 0.787  | 0.772 | 0.841 | 0.109 | 0.904 | 0.900 | 0.939 | 0.049 | 0.874 | 0.868 | 0.907 | 0.071 |
| V  | +MAC+NL+CMD | 0.797  | 0.785 | 0.853 | 0.107 | 0.905 | 0.902 | 0.939 | 0.048 | 0.880 | 0.881 | 0.917 | 0.066 |
| VI | +SMAC+NL+SCMD | 0.801  | 0.790 | 0.857 | 0.098 | 0.911 | 0.908 | 0.942 | 0.043 | 0.878 | 0.874 | 0.909 | 0.064 |

Gradually Using Proposed Model Components

Compare Row I and II with Using Self-Attention [21]

| VII | +SA [21] | 0.765 | 0.749 | 0.823 | 0.113 | 0.886 | 0.878 | 0.925 | 0.054 | 0.821 | 0.796 | 0.864 | 0.093 |

Compare Row IV with Using $S^2$MA [32]

| VIII | $S^2$MA [32] | 0.786 | 0.770 | 0.841 | 0.110 | 0.901 | 0.896 | 0.936 | 0.050 | 0.862 | 0.862 | 0.906 | 0.076 |

Compare Row V with Using the Concatenation Based CMD in All Five Decoders

| IX  | +MAC+NL+CMD | 0.787 | 0.774 | 0.844 | 0.110 | 0.908 | 0.908 | 0.943 | 0.046 | 0.876 | 0.881 | 0.916 | 0.066 |

Compare Row VI with Using Naive Fusion Methods

![Fig. 9. Visualization of some attention maps and feature maps.](image)

We show the feature maps ($X^r$), the attention maps ($A^r(X^r)$), the contrast attention maps ($C^r(X^r)$), and the output feature maps of the SMAC module ($Z^r$) for the RGB and depth modalities in two image pairs. In each image, the white point indicates the query position.

![Fig. 10. Example depth maps for the smallest (top row) and largest (bottom row) selective attention weights.](image)

DenseASPP for the RGB data only. In the second and the third row we show the model performance of using MA and MAC modules, respectively, to incorporate attention-based cross-modal interaction. We can see that adopting MA largely improves the model performance, especially on the LFSD dataset. The MAC module can further moderately improve the results by incorporating the contrast mechanism on two out of the three datasets. Then, we add NL modules on top of the output feature maps of the MAC module to further blend the received cross-modal cues. The results are reported in row IV and we find that further using NL modules after the MAC module can continue bringing performance gains, especially on the LFSD dataset. Next, we use cross-modal decoders (CMD) by adopting MA modules in the first three decoders and the simple concatenation based fusion method in the last two, as discussed in Section 5. The results in row V demonstrate that fusing cross-modal features in decoder modules can also promote the modal capability, especially on the RedWeb-S dataset. At last, we adopt the computed selective attention in MAC and CMD to weight the depth cues and report the results in row VI. We observe that using this strategy can lead to performance improvements on two out of the three datasets. Hence, we use this model setting as our final RGB-D SOD model.

To thoroughly understand the effectiveness of our proposed SMAC module, we show some visualization examples of the input feature maps $X^r$, the attention maps $A^r(X^r)$, the contrast attention maps $C^r(X^r)$, and the output feature maps $Z^r$ of the SMAC module in Figure 9, for both RGB and depth modalities. In the left example, the RGB image is more discriminative than the depth map, thus the RGB attention maps $A^r(X^r)$ and $C^r(X^r)$ are better than the depth attention maps. An opposite situation is given in the right example. We can see that after adopting the SMAC module, the four feature maps are all improved and become...
and cascade NL right after the MAC module.

To evaluate the necessity of adopting MA modules in decoders, we report the results of using the concatenation based CMD in all five decoders in row IX. Compared with the results in row V, we observe that substituting the concatenation based fusion with MA in the first three decoders achieves large performance gains on two out of the three datasets, i.e., ReDWeb-S and LFSFD. This result demonstrates the superiority of adopting the MA mechanism on multi-level features.

Finally, we also compare with some naive fusion methods, such as concatenation, summation, and multiplication. We adopt them after the encoder and in every decoder module and report their results in rows X to XII. We can see that our final results in row VI outperform theirs on all the three datasets by a large margin, thus demonstrating the necessity of adopting high-order cross-modal information interactions.

### 6.4 Comparison with State-of-the-Art Methods

We compare our final SMAC RGB-D SOD model with state-of-the-art methods on all the nine datasets. Specifically, the compared methods include DF [7], AFNet [12], CTMF [16], MMCI [14],
PCF [17], TANet [20], CPFP [36], DMRA [13], D3Net [9], and our previous $S^2$MA [32]. To demonstrate the benefit of our proposed ReDWeb-S dataset, we also train a SMAC model additionally using its training set and name this model as SMAC*.

The quantitative comparison results are given in Table 3. From the comparison we observe that our new SMAC model outperforms previous methods on eight datasets. Especially, it outperforms our previous $S^2$MA SOD model by a large margin on most datasets, demonstrating the effectiveness of our extensions. Comparing SMAC* with SMAC, we can conclude that including ReDWeb-S in the training set can improve the results on most datasets, especially on SIP, which mainly contains real-world scenes. However, on RGBD135 and DUTLF-Depth, SMAC* is worse than SMAC. This is reasonable since these two datasets both mainly focus on simplex artifacts in a close range. These observations verify the benefit of our proposed ReDWeb-S dataset for real-word RGB-D SOD.

We also show the qualitative comparison of some ReDWeb-S images in Figure 11. The first three rows show three images with very complex visual scenes, such as complex foreground objects and cluttered backgrounds. We can see that such very challenging scenarios are very difficult for most previous methods while our SMAC and SMAC* models can successfully locate most parts of the salient objects. Furthermore, we also show images with small salient objects, large salient objects, multiple salient objects, faraway salient objects, and inconspicuous salient objects in rows from IV to VIII, respectively. Our models can handle all these challenging situations and outperform previous methods, demonstrating the effectiveness of our proposed SMAC model.

6.5 Effectiveness of the ReDWeb-S Dataset

In this section, we further analyze the effectiveness of the proposed ReDWeb-S dataset by comparing the model performance of using different training sets. We first train our SMAC SOD model using one of the four training datasets, i.e., NJUD, NLPR, DUTLF-Depth, and ReDWeb-S. Due to the space limitation, we do not report all the results on the nine datasets in terms of the four metrics. Instead, we follow [9] to rank the performance of different training settings on each dataset under each evaluation metric. Finally, we compute and report the average performance rank for each setting on all the nine datasets and using the four metrics. We also evaluate the model performance of using every two of the four datasets as the training set. Table 4 shows the results. We observe that when only using one dataset as the training set, ReDWeb-S performs not good. This is because its real-world scenes have large differences with many datasets. However, when training using two datasets, combining ReDWeb-S with NJUD achieves
the best performance among all the six training settings, which
contributes the complementary role of our proposed dataset for
existing ones.

7 Conclusion
In this paper, we first propose a real-world RGB-D SOD bench-
mark dataset. Different from the numerous previous datasets, it
has the most image pairs with high-quality depth maps, and the
most diverse visual scenes and objects. Hence, it has both high-
quality and challenging, which bring large benefits for both model
training and comprehensive model evaluation. For a new and
more effective way to fuse cross-modal information for RGB-D
SOD, we propose a novel mutual attention model to fuse non-
local attention and context features from different modalities,
and also achieve high-order and trilinear modality interaction. We
also incorporate the contrast mechanism and obtain a unified model.
A selective attention mechanism is also presented to reweight the
depth cues thus reducing the impact of low-quality depth data.
By embedding the proposed SMAC model into a two-stream
UNet architecture, we outperform state-of-the-art RGB-D SOD
methods. We also thoroughly analyze the effectiveness of the
SMAC model and the proposed dataset.

References
[1] N. Liu and J. Han, “Dhsnet: Deep hierarchical saliency network for
saliency object detection,” in CVPR, 2016, pp. 678–686.
[2] N. Liu, J. Han, and M.-H. Yang, “Picnet: Pixel-wise contextual attention
learning for accurate saliency detection,” IEEE Transactions on Image
Processing, 2020.
[3] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, “An iterative and
cooporative top-down and bottom-up inference network for salient object
detection,” in Proceedings of the IEEE Conference on Computer Vision
and Pattern Recognition, 2019, pp. 5968–5977.
[4] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, “Eg-
net: Edge guidance network for salient object detection,” in Proceedings
of the IEEE International Conference on Computer Vision, 2019, pp.
8779–8788.
[5] X. Qiu, Z. Zhang, C. Huang, C. Gao, M. Dehghani, and M. Jagersand,
“Basnet: Boundary-aware salient object detection,” in Proceedings of the
IEEE Conference on Computer Vision and Pattern Recognition, 2019,
pp. 7479–7489.
[6] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, “Depth-aware
saliency object detection and segmentation via multiscale discriminative
saliency fusion and bootstrap learning,” TIP, vol. 26, no. 9, pp. 4204–
4216, 2017.
[7] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, “Rgbd salient
object detection via deep fusion,” TIP, vol. 26, no. 5, pp. 2274–2285,
2017.
[8] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, “Saliency object
detection for rgb-d image by single stream recurrent convolution neural
network,” Neurocomputing, vol. 363, pp. 46–57, 2019.
[9] D.-P. Fan, Z. Lin, J.-X. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, and M.-
M. Cheng, “Rethinking rgb-d salient object detection: Models, datasets,
and large-scale benchmarks,” arXiv preprint arXiv:1907.06781, 2019.
[10] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, “Depth really
matters: Improving visual salient region detection with depth,” in BMVC,
2013.
[11] J. Guo, T. Ren, and J. Bei, “Saliency object detection for rgb-d image via
saliency evolution,” in ICME. IEEE, 2016, pp. 1–6.
[12] N. Wang and X. Gong, “Adaptive fusion for rgb-d salient object detec-
tion,” IEEE Access, vol. 7, pp. 55277–55284, 2019.
[13] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, “Depth-induced multi-scale
recurrent attention network for saliency detection,” in ICCV, 2019.
[14] H. Chen, Y. Li, and D. Su, “Multi-modal fusion network with multi-
scale multi-path and cross-modal interactions for rgb-d salient object
detection,” Pattern Recognition, vol. 86, pp. 376–385, 2019.
[15] R. Shigematsu, D. Feng, S. You, and N. Barnes, “Learning rgb-d salient
object detection using background enclosure, depth contrast, and top-
down features,” in ICCV Workshops, 2017, pp. 2749–2757.
[45] Z. Zhang, “Microsoft kinect sensor and its effect,” *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[46] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in *CVPR*, 2017, pp. 2462–2470.

[47] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, “Monocular relative depth perception with web stereo data supervision,” in *CVPR*, 2018, pp. 311–320.

[48] G. Lin, A. Milan, C. Shen, and I. Reid, “Refinenet: Multi-path refinement networks for high-resolution semantic segmentation,” in *CVPR*, 2017, pp. 1925–1934.

[49] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.

[50] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[51] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” pp. 1492–1500, 2017.

[52] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, “Fixing the train-test resolution discrepancy,” pp. 8252–8262, 2019.

[53] J. Deng, W. Dong, R. Socher, L. J. Li, and F. F. Li, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, 2009, pp. 248–255.

[54] S. Xiang, L. Yu, and C. W. Chen, “No-reference depth assessment based on edge misalignment errors for rs-d images,” *TPP*, vol. 25, no. 3, pp. 1479–1494, 2015.

[55] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, “Salient objects in clutter: Bringing salient object detection to the foreground,” in *ECCV*, 2018, pp. 186–202.

[56] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, “Salient object detection: A benchmark,” *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[57] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, “Interactive two-stream decoder for accurate and fast saliency detection,” in *CVPR*, 2020, pp. 9141–9151.

[58] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *ICML*, 2015, pp. 448–456.

[59] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *ICLR*, 2015.

[60] N. Liu, J. Han, and M.-H. Yang, “Picanet: Learning pixel-wise contextual attention for saliency detection,” in *CVPR*, 2018, pp. 3089–3098.

[61] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[62] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015, pp. 3431–3440.

[63] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, “Denseasc for semantic segmentation in street scenes,” in *CVPR*, 2018, pp. 3684–3692.

[64] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[65] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in *ICCV*, 2017, pp. 4558–4567.

[66] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” in *IJCAI*. AAAI Press, 2018, pp. 698–704.

[67] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, “2adele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection,” in *CVPR*, 2020, pp. 9060–9069.

[68] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in *NIPS Autodiff Workshop*, 2017.