Learning Selective Self-Mutual Attention for RGB-D Saliency Detection

Nian Liu[1,2] Ni Zhang[1] Junwei Han[1*]
[1]Northwestern Polytechincal University [2]Mohamed bin Zayed University of Artificial Intelligence
{liunian228, nnizhang.1995, junweihan2010}@gmail.com

Abstract

Saliency detection on RGB-D images is receiving more and more research interests recently. Previous models adopt the early fusion or the result fusion scheme to fuse the input RGB and depth data or their saliency maps, which incur the problem of distribution gap or information loss. Some other models use the feature fusion scheme but are limited by the linear feature fusion methods. In this paper, we propose to fuse attention learned in both modalities. Inspired by the Non-local model, we integrate the self-attention and each other's attention to propagate long-range contextual dependencies, thus incorporating multi-modal information to learn attention and propagate contexts more accurately. Considering the reliability of the other modality's attention, we further propose a selection attention to weight the newly added attention term. We embed the proposed attention module in a two-stream CNN for RGB-D saliency detection. Furthermore, we also propose a residual fusion module to fuse the depth decoder features into the RGB stream. Experimental results on seven benchmark datasets demonstrate the effectiveness of the proposed model components and our final saliency model. Our code and saliency maps are available at https://github.com/nnizhang/S2MA.

1. Introduction

Saliency detection is the task to distinguish the most salient object from complex background in a visual scene. It mimics the human visual attention mechanism to find out what catch people's eyes when free viewing scenes. This task can be used as a pre-processing technique or a model component for many other vision tasks, such as semantic segmentation [52, 2], and image editing [43, 50].

Researchers have proposed many saliency models in decades, and most of them work on RGB images. Although recent RGB saliency models have achieved very promising performance, *e.g.*, [32, 33, 17, 31], they can only leverage appearance cues from the input RGB data, which are often severely limited in many challenging scenarios, *e.g.*, cluttered background or salient object having similar appearance with the background. Nevertheless, we human beings actually live in a 3D environment, in which depth cues can supply sufficient complementary information for the appearance cues. Thus, it is quite necessary to study the saliency detection problem on RGB-D data. In Figure 1 we show a challenging image with complex appearance. In column (b), we can see that its RGB-based saliency detection result is easily disturbed and has severe false positive highlights. However, its foreground object is quite different from the background in terms of depth. Thus the depth information can be used to easily distinguish the foreground object and obtain an accurate saliency map.

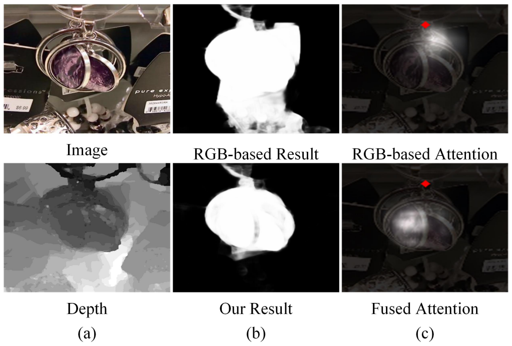


Figure 1. Comparison of the RGB and RGB-D saliency detection results and the learned attention. In (a) we give a challenging image and its corresponding depth map. (b) shows the saliency maps of a RGB-based deep model (top) and our RGB-D-based model (bottom). (c) illustrates the learned attention maps at the same position (the red point) in a RGB-based Non-local model [51] (top) and our proposed attention fusion model (bottom).

To combine the appearance information and the depth cues for RGB-D salient object detection, some previous methods adopt the **early fusion** strategy [46, 41, 34, 14], in which case both RGB and depth data are taken as inputs and processed in a unified model. However, it is not easy for one model to fit the data from two modalities well due to their distribution gap. Some other models use the **result fusion** strategy [11, 22, 49], where the RGB image and the depth

[*] Corresponding author

map are used in two models to generate their own saliency maps separately, and then a fusion method is adopted to fuse the two saliency maps. This scheme is also suboptimal since rich modality information is gradually compressed and lost in the two separate saliency modeling processes. Thus, the final interaction between the two saliency maps is highly limited.

As a better choice, many models utilize the **middle fusion** strategy, i.e., fusing the intermediate information of the two modalities, and then generating the final saliency map. Most typically, many recently proposed deep RGB-D saliency models [40, 5, 44, 23, 3, 4, 19] first use two-stream CNNs to extract RGB and depth features separately, and then fuse them via summation or concatenation. We refer these methods as the **feature fusion** strategy. This strategy avoids the distribution gap problem and fuses rich multi-modal features with plentiful interaction. However, simple feature summation or concatenation only learns to linearly fuse RGB and depth features, being unable to explore more complex multi-modal interaction.

In this paper, we present a novel middle fusion strategy. Inspired by the Non-local (NL) model [51], we propose to fuse multi-modal attention. The NL model computes each position a set of spatial attention and then uses them to aggregate the features at all positions, thus being able to incorporate long-range global context. Since the attention and the propagated features in the NL model are based on the same feature map, this kind of attention mechanism is usually referred as the **self-attention**. Considering the complementary information between the RGB and the depth data, we propose to further propagate global context using each other's attention, as which we refer the **mutual-attention** mechanism. Since the original self-attention may be highly limited by the single modality information, the proposed mutual attention can supply extra complementary cues about where should attend based on the information of the other modality when propagating the contextual features. In Figure 1(c), we show two attention maps of the same position (the red point) in the image. We can see that RGB-based attention is highly biased due to the complex appearance, while the fused attention can accurately locate the main body of the salient object by fusing the depth attention.

Furthermore, since the complementary information from the other modality may be not always reliable for all positions, we propose another selection attention to decide how much mutual attention should be involved at each position. We adopt this novel selective self-mutual attention mechanism in a two-stream CNN network to fuse multi-modal cues for RGB-D saliency detection. Additionally, we also present a novel residual fusion module to transfer the depth cues to the RGB features in the decoder part. Experimental results demonstrate that our proposed model components are all helpful for improving the saliency detection performance. Finally, our saliency model outperforms all other state-of-the-art methods.

2. Related Work

Saliency detection on RGB-D images. Early RGB-D salient object detection methods usually borrow common priors (*e.g.*, contrast [8] and compactness [10]) from RGB saliency models to design RGB and depth features. Additionally, some researchers exploit depth-specific priors, *e.g.*, shape and 3D layout priors [9], and anisotropic center-surround difference [28, 22].

Recent work introduce CNNs to RGB-D salient object detection and have achieved promising results. Qu *et al.* [41] adopt the early fusion strategy to serialize hand-crafted RGB and depth features together as the CNN inputs. Fan *et al.* [14] and Liu *et al.* [34] concatenate each depth map as the 4^{th} channel with the corresponding RGB image as the CNN input. Wang *et al.* [49] adopt the result fusion strategy and learn a switch map to adaptively fuse RGB and depth saliency maps. Many recent work adopt the middle fusion strategy to fuse intermediate depth and appearance features. Han *et al.* [23] fuse the representation layers of the RGB and the depth modality with a joint representation layer. Chen *et al.* [3] propose a complementarity-aware fusion module to capture cross-modal and cross-level features. In contrast, our model focuses on fusing multi-modal attention.

Some other existing models also leverage the attention mechanism to fuse the two modalities, *e.g.*, [55] and [40]. However, they only generate channel [40] or spatial [55] attention from the depth view and adopt it to filter the appearance features. Nevertheless, we generate non-local attention from both views and then fuse them to propagate long-range contexts.

Non-local networks. Vaswani *et al.* [47] propose a self-attention network for language modeling. Given a query and a set of key-value pairs, they first compute the attention weight between the query and each key. Then they use the attention weights to aggregate the values by weighted sum as the output. Similarly, Wang *et al.* [51] propose the NL model for learning self-attention in 2D or 3D vision modeling. Huang *et al.* [26] propose to replace the densely connected attention path with the criss-cross path to improve the model efficiency. Cao *et al.* [1] propose to unify the NL model with SENet [24] to learn query-independent global context with a lightweight network structure. In [18], Fu *et al.* apply the NL model to capture both spatial and channel long-range dependencies. Some other models [7, 30, 57, 54] propose to improve both model performance and efficiency by learning representative key-value pairs. In this paper, we propose to improve the NL module by fusing attention from multi-modalities, thus greatly promoting the accuracy of attention generation and context propagation.

Multi-modal attention learning. In [36] and [48], the authors also propose to learn multi-modal attention. Nam *et al*. [36] propose to learn visual and textual attention mechanisms for both multi-modal reasoning and matching. Wan *et al*. [48] apply 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. On the contrary, we propose to directly fuse multi-modal attention.

3. From Non-local to Selective Self-Mutual Attention

In this section, we elaborate on the proposed Selective Self-Mutual Attention (S^2MA) module for fusing multi-modal information. It is built on the basis of the NL module [51], additionally with the proposed attention fusion and selective attention added. We first briefly review the NL module and then go into our S^2MA module.

3.1. Reviewing the NL module

Here we briefly review the NL module, whose network architecture is shown in Figure 2(a). Imaging we have a feature map $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent its height, width, and channel number, respectively, the NL module first embeds $\boldsymbol{X}$ into three spaces with C_1 channels:

$$\theta(\boldsymbol{X}) = \boldsymbol{X}W_\theta, \phi(\boldsymbol{X}) = \boldsymbol{X}W_\phi, g(\boldsymbol{X}) = \boldsymbol{X}W_g, \quad (1)$$

where W_θ, W_ϕ, and $W_g \in \mathbb{R}^{C \times C_1}$ are embedding weights, and the embeddings can be implemented by 1×1 convolution as shown in Figure 2(a).

Next, a similarity (or affinity) function f is computed within θ and ϕ embeddings. In [51], several forms of the function f are proposed. Here we introduce the most widely used embedded Gaussian function, where

$$f(\boldsymbol{X}) = \theta(\boldsymbol{X})\phi(\boldsymbol{X})^\top, \quad (2)$$

and $f(\boldsymbol{X}) \in \mathbb{R}^{HW \times HW}$. In $f(\boldsymbol{X})$, each element $f_{i,j}$ represents the affinity between the i^{th} and the j^{th} spatial location in $\boldsymbol{X}$.

Subsequently, the NL module generates the attention weight by using normalization along the second dimension:

$$A(\boldsymbol{X}) = softmax(f(\boldsymbol{X})), \quad (3)$$

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

$$\boldsymbol{Y} = A(\boldsymbol{X})g(\boldsymbol{X}), \quad (4)$$

where $\boldsymbol{Y} \in \mathbb{R}^{HW \times C_1}$ is an attentive feature, and it is further reshaped to the shape $H \times W \times C_1$.

Finally, the NL module learns a residual signal based on $\boldsymbol{Y}$ to improve the original feature $\boldsymbol{X}$ and obtain the final output $\boldsymbol{Z}$:

$$\boldsymbol{Z} = \boldsymbol{Y}W_Z + \boldsymbol{X}, \quad (5)$$

where $W_Z \in \mathbb{R}^{C_1 \times C}$ is the weight of a 1×1 Conv layer for projecting the attentive feature back to the original feature space.

3.2. Self-Mutual Attention

The obtaining of the attention $A(\boldsymbol{X})$ in the NL module can be rewritten as:

$$A(\boldsymbol{X}) = softmax(\boldsymbol{X}W_\theta W_\phi^\top \boldsymbol{X}^\top). \quad (6)$$

We can see that it is a bilinear projection of the original feature $\boldsymbol{X}$ itself, thus the NL module belongs to the self-attention category. We argue that using a further projection of the same feature can only bring limited information and performance gain (see the experimental results in Section 5.4). For multi-modal tasks, such as RGB-D salient object detection, we can leverage the features from multiple modalities to integrate information complementarity.

In this paper, we first propose fusing Self-Mutual Attention (SMA) to improve the NL module for multi-modal data. Considering we have two feature maps $\boldsymbol{X}^r, \boldsymbol{X}^d \in \mathbb{R}^{H \times W \times C}$ from the RGB and the depth modality, respectively, we follow the NL module to embed them into the θ, ϕ, g spaces and obtain their affinity matrixes respectively:

$$\begin{aligned} f^r(\boldsymbol{X}^r) &= \theta^r(\boldsymbol{X}^r)\phi^r(\boldsymbol{X}^r)^\top, \\ f^d(\boldsymbol{X}^d) &= \theta^d(\boldsymbol{X}^d)\phi^d(\boldsymbol{X}^d)^\top. \end{aligned} \quad (7)$$

Since the two affinity matrixes are computed by their own modality-specific feature, we fuse them via simple summation and then obtain the fused attention:

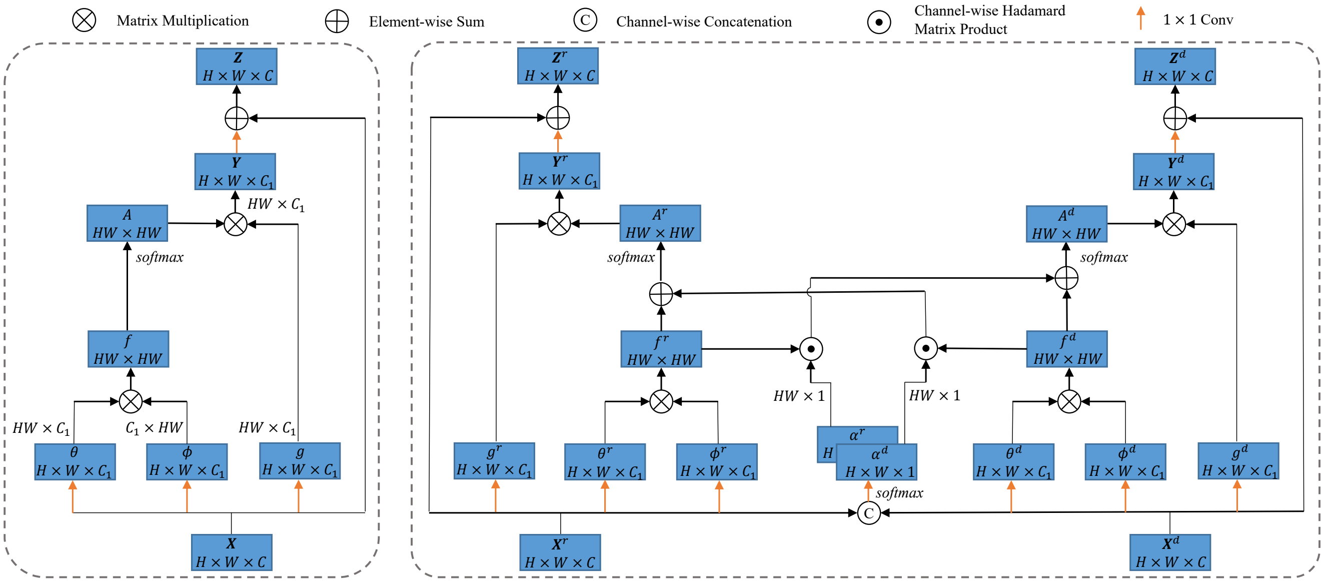
$$A^f(\boldsymbol{X}^r, \boldsymbol{X}^d) = softmax(f^r(\boldsymbol{X}^r) + f^d(\boldsymbol{X}^d)). \quad (8)$$

Then we use A^f to propagate long-range contextual dependencies in the two modalities, respectively:

$$\begin{aligned} \boldsymbol{Y}^r &= A^f(\boldsymbol{X}^r, \boldsymbol{X}^d)g^r(\boldsymbol{X}^r), \\ \boldsymbol{Y}^d &= A^f(\boldsymbol{X}^r, \boldsymbol{X}^d)g^d(\boldsymbol{X}^d). \end{aligned} \quad (9)$$

Finally, we use (5) to obtain modality-specific outputs $\boldsymbol{Z}^r$ and $\boldsymbol{Z}^d$, respectively.

Note that in A^f, the affinity of two modalities are both included, thus the attention generation and context propagation can become more accurate. Experimental results in Section 5.4 demonstrate that the SMA module can bring significant performance gain for the RGB-D saliency detection task.



(a) Non-local Module

(b) Selective Self-Mutual Attention Module

Figure 2. Network architecture of the original Non-local module [51] (a) and the proposed selective self-mutual attention module (b).

3.3. Selective Self-Mutual Attention

The SMA model considers self-mutual attention equally. However, the mutual attention from the other modality is not always reliable for all positions since the information from one modality may be inaccurate or useless for some positions. A typical example is that some depth maps are noisy and inaccurate in some datasets. Based on the widely validated effectiveness of the self-attention mechanism and the experimental results, we choose to reweight the mutual attention by computing a selection attention weight at each position. Specifically, we first concatenate $\boldsymbol{X}^r$ and $\boldsymbol{X}^d$ and then use a 1×1 Conv layer with the softmax activation function to compute the selection attention:

$$\alpha = softmax(Conv([\boldsymbol{X}^r, \boldsymbol{X}^d])), \tag{10}$$

where $\alpha \in \mathbb{R}^{H \times W \times 2}$ and $[\cdot]$ indicates the concatenation operation. We further split it into two maps $\alpha^r, \alpha^d \in \mathbb{R}^{H \times W \times 1}$. Each of them represents the reliability at all positions of the corresponding modality.

Then, we can obtain the selective self-mutual attention by changing (8) with a weighted sum of the two affinities:

$$\begin{aligned} A^r(\boldsymbol{X}^r, \boldsymbol{X}^d) &= softmax(f^r(\boldsymbol{X}^r) + \alpha^d \odot f^d(\boldsymbol{X}^d)), \\ A^d(\boldsymbol{X}^r, \boldsymbol{X}^d) &= softmax(f^d(\boldsymbol{X}^d) + \alpha^r \odot f^r(\boldsymbol{X}^r)), \end{aligned} \tag{11}$$

where $\odot$ is the channel-wise Hadamard matrix product. Finally, we use A^r and A^d to aggregate contextual features for the two views, respectively, similar with (9).

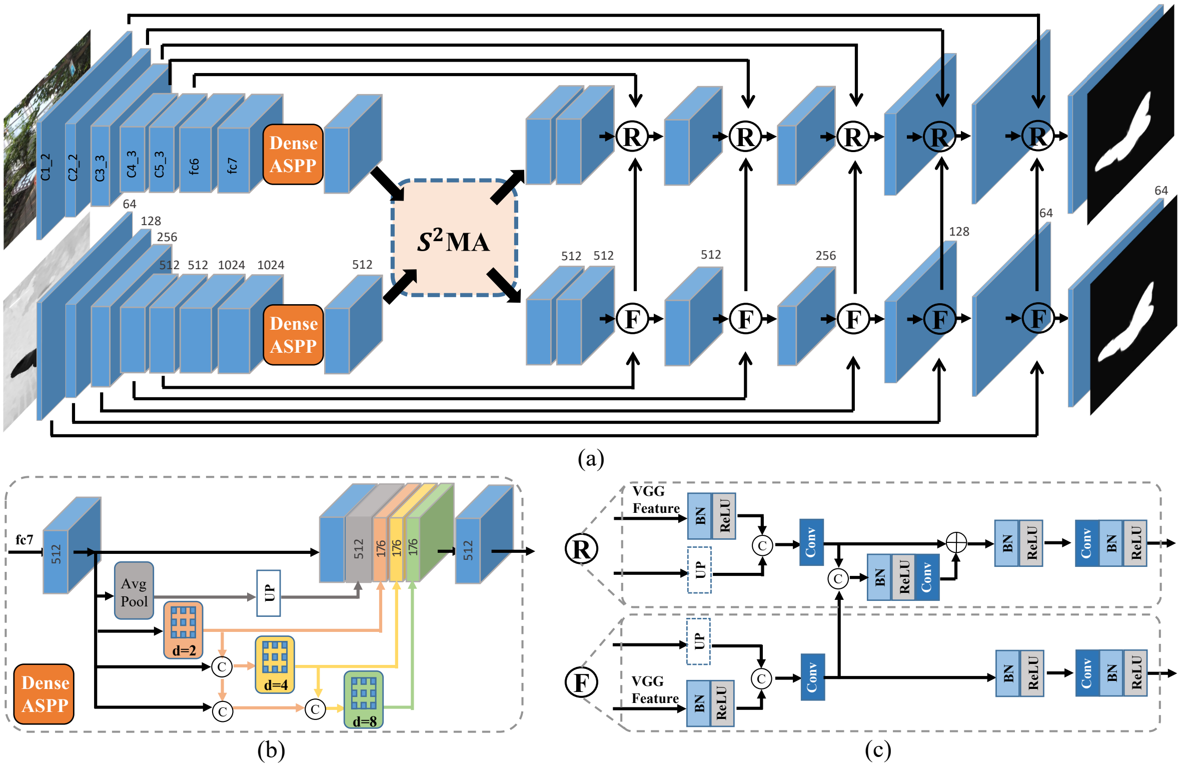
The whole network architecture of the S^2MA module is shown in Figure 2(b). Experimental results in Section 5.4 indicate that using the proposed selection attention can further improve the model performance on the basis of the SMA module.

4. RGB-D Saliency Detection Network

Based on the proposed S^2MA module, we propose a novel deep model for RGB-D saliency detection. As shown in Figure 3(a), our model is based on a two-stream CNN, and each of them is based on a UNet [42] architecture.

Specifically, we take the VGG-16 network [45] as the backbones of the UNets and share the same network structure for the two encoder parts. We follow [33] to slightly change the VGG-16 network structure as follows. First, we change the pooling strides of the pool4 and pool5 layers to 1 and the dilation rates [6] of the conv5 block to 2. Second, we turn the fc6 layer into a 3×3 Conv layer with 1024 channels and the dilation rate of 12. Third, we transform the fc7 layer to a 1×1 Conv layer with 1024 channels. As such, the encoder network becomes a fully convolutional network [35] with the output stride of 8.

Next, to further enhance the capability of the encoder network, we adopt a DenseASPP [53] module, which introduces dense connections [25] to the ASPP [6] module to cover dense feature scales. We first adopt a 1×1 Conv layer to compress the fc7 feature map to 512 channels and then deploy the DenseASPP module on it. Considering our specific training image size, we design three dilated Conv branches, which have 3×3 Conv layers with 176 channels and the dilation rates of 2, 4, and 8, respectively. At the same time, we follow [53] to densely connect the three branches. To capture the global feature, we also design another branch to average pool the input feature map and then



(a)

(b)

(c)

Figure 3. Architecture of our proposed RGB-D saliency detection 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. (b) shows our used DenseASPP module. Some key channel numbers are also given. (c) shows the two proposed decoder modules of the RGB and the depth CNN streams, respectively. Here "UP" means upsampling with bilinear interpolation.

upsample it to the original size. Finally, we concatenate the input feature map and the features of all the four branches, and then compress them to 512 channels. The whole module architecture can be found in Figure 3(b).

After the DenseASPP module, we take the features of the RGB and the depth CNN streams as the inputs and adopt the proposed S^2MA module to fuse multi-modal attention and propagate global contexts for both views. Whereafter, we go into the decoding part. In the first decoder module, we use two Conv layers with 512 channels. Then we follow the UNet [42] architecture to progressively skip connect intermediate encoder features with decoder features. The used intermediate VGG features are the last Conv feature maps of the five blocks, which are marked in Figure 3(a). For each depth decoder module, we use a naive fusion model Ⓕ by simply concatenating the VGG feature and the previous decoder feature, and then adopting two Conv layers to fuse them. For the RGB decoder modules, we design a residual fusion model Ⓡ to further fuse depth decoder features with a residual connection. Specifically, after concatenating the two input features and adopt the first Conv layer, we concatenate this feature with the first Conv feature of the corresponding depth decoder module as the pre-activation of the fused feature. Then we use another Conv layer to learn a residual fusion signal to ease the network training. The detailed network structure is shown in Figure 3(c). Please note that we do not further adopt the S^2MA module in the decoding part since it is computational prohibitive for large feature maps.

Each Conv layer in our decoder part has 3×3 kernels, and is followed by a BN [27] layer and the ReLU activation function. The output channel number in each decoder module is set to be the same as that of the next skip-connected VGG feature, which is also marked in Figure 3(a). For each of the last three decoder modules, since the previous decoder feature map has a smaller spatial size than the skip-connected VGG feature map, we upsample it by bilinear interpolation to progressively enlarge the spatial size. Finally, we adopt a 3×3 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.

5. Experiments

5.1. Datasets

For model training and evaluation, we use seven RGB-D saliency benchmark datasets as follows. **NJUD** [28] has 1,985 images collected from the Internet, 3D movies, and

Table 1. Ablation study on the effectiveness of the DenseASPP (DA) module, the NL module, the SMA module, the S^2MA module, and the residual fusion module ®. **Blue** indicates the best performance.

Settings					NJUD [28]				NLPR [39]				RGBD135 [8]				LFSD [29]			
DA	NL	SMA	S^2MA	®	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
					0.865	0.852	0.902	0.072	0.897	0.873	0.941	0.039	0.875	0.834	0.927	0.046	0.786	0.775	0.836	0.131
✓					0.877	0.865	0.913	0.057	0.911	0.892	0.945	**0.030**	0.889	0.861	0.925	0.032	0.787	0.768	0.836	0.118
✓	✓				0.877	0.865	0.916	0.057	0.908	0.888	0.945	0.032	0.892	0.868	0.922	0.034	0.793	0.784	0.838	0.123
✓		✓			0.890	0.882	0.927	0.058	0.907	0.886	0.947	0.035	0.918	0.903	0.956	0.028	0.821	0.812	0.857	0.108
✓			✓		0.889	0.884	0.929	0.056	**0.915**	0.898	0.950	0.031	0.929	0.918	0.972	0.025	0.829	0.819	0.865	0.101
✓			✓	✓	**0.894**	**0.889**	**0.930**	**0.053**	**0.915**	**0.902**	**0.953**	**0.030**	**0.941**	**0.935**	**0.973**	**0.021**	**0.837**	**0.835**	**0.873**	**0.094**

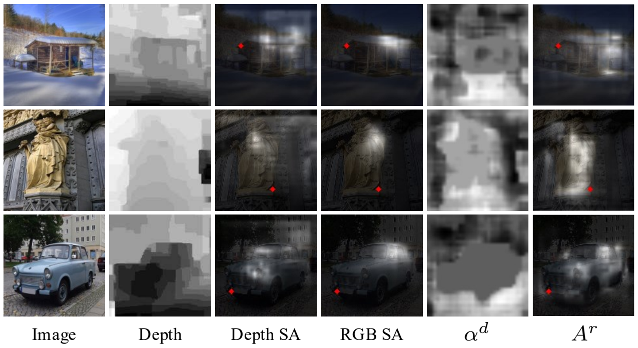


Figure 4. Visualization of the learned attention maps. We show the learned depth-based self-attention (Depth SA), the RGB-based self-attention (RGB SA), the selection attention α^d, and the fused attention A^r in three images. In each image, the red point indicates the query position.

stereo photos. **NLPR** [39] and **RGBD135** [8] contain 1,000 and 135 images collected by the Microsoft Kinect, respectively. **LFSD** [29] contains 100 images captured by a Lytro light field camera. **STERE** [37] contains 1,000 pairs of binocular images downloaded from the Internet. **SSD** [56] has 80 stereo movie frames. **DUT-RGBD** [40] contains 1,200 real life images captured by a Lytro2 camera.

5.2. Evaluation Metrics

Following recent work, we adopt four evaluation metrics. The first one is the maximum F-measure (maxF). F-measure comprehensively considers both precision and recall for binarized saliency maps and we report the maxF score under an optimal threshold. The second metric is the Structure-measure S_m [12] which evaluates both region-aware and object-aware structural similarities between the saliency maps and the ground truth. We use the third metric as the Enhanced-alignment measure E_ξ [13] to capture both global statistics and local pixel matching information of the saliency maps. The fourth metric is the Mean Absolute Error (MAE). It measures the average of the per-pixel absolute difference between the saliency maps and the ground truth.

5.3. Implementation Details

For fair comparisons, we adopt the same training set as in [23, 3, 55], which consists of 1,400 images from the NJUD dataset and 650 images from the NLPR dataset. For data augmentation, we first resize training images and corresponding depth maps to 288 × 288 pixels and then randomly crop 256 × 256 image regions to train the network. Random horizontal flipping is also used. For the depth stream CNN, we simply replicate each single depth map to three channels to fit the network input layer. Since the depth maps of different datasets have different presentations, we process them to a unified presentation, where small depth values represent the object is close to the camera and vice verse. We also normalize the depth map of each image to the value range of [0,255] to ease the network training. Finally, each image and the three-channel depth map are subtracted by their mean pixel values as the inputs of the two-stream network.

We use the cross-entropy loss between the predicted saliency maps and the ground truth masks as the training losses of both streams. To facilitate the network training, we also use deep supervision for each decoder module, where we first adopt a 3 × 3 Conv layer with the sigmoid activation function on each decoder feature map to generate a saliency map and then compute the cross entropy loss to train this decoder module. Following [33], we empirically use 0.5, 0.5, 0.5, 0.8, and 0.8 to weight the first five decoder losses of each stream. The stochastic gradient descent (SGD) with momentum algorithm is adopted to optimize our saliency network with totally 40,000 iterations. Weight decay, momentum, and batchsize are set to 0.0005, 0.9, and 8, respectively. The initial learning rate is set to 0.01 and divided by 10 at the $20,000^{th}$ and the $30,000^{th}$ iteration, respectively.

We implement the proposed network using the Pytorch [38] package and use a GTX 1080 Ti GPU for computing acceleration. During testing, we directly resize each image and its depth map to 256 × 256 pixels as the network inputs and obtain the saliency map from the network output of the RGB stream without any post-processing method. The testing process takes 0.107 seconds for each image.

5.4. Ablation Study

To evaluate the effectiveness of the proposed model components, we conduct the ablation study on four datasets,

Table 2. Quantitative comparison of our proposed model with other 11 state-of-the-art RGB-D saliency models on 7 benchmark datasets in terms of 4 evaluation metrics. **Red** and **blue** indicate the best and the second best performance, respectively.

Dataset	Metric	LBE [16]	DCMC [10]	SE [22]	DF [41]	AFNet [49]	CTMF [23]	MMCI [5]	PCF [3]	TANet [4]	CPFP [55]	DMRA [40]	S^2MA (Ours)
NJUD	S_m ↑	0.695	0.686	0.664	0.763	0.772	0.849	0.858	0.877	0.878	0.878	**0.886**	**0.894**
	maxF ↑	0.748	0.715	0.748	0.804	0.775	0.845	0.852	0.872	0.874	0.877	**0.886**	**0.889**
	E_ξ ↑	0.803	0.799	0.813	0.864	0.853	0.913	0.915	0.924	0.925	0.923	**0.927**	**0.930**
[28]	MAE ↓	0.153	0.172	0.169	0.141	0.100	0.085	0.079	0.059	0.060	**0.053**	**0.051**	**0.053**
NLPR	S_m ↑	0.762	0.724	0.756	0.802	0.799	0.860	0.856	0.874	0.886	0.888	**0.899**	**0.915**
	maxF ↑	0.745	0.648	0.713	0.778	0.771	0.825	0.815	0.841	0.863	0.867	**0.879**	**0.902**
	E_ξ ↑	0.855	0.793	0.847	0.880	0.879	0.929	0.913	0.925	0.941	0.932	**0.947**	**0.953**
[39]	MAE ↓	0.081	0.117	0.091	0.085	0.058	0.056	0.059	0.044	0.041	0.036	**0.031**	**0.030**
RGBD135	S_m ↑	0.703	0.707	0.741	0.752	0.770	0.863	0.848	0.842	0.858	0.872	**0.900**	**0.941**
	maxF ↑	0.788	0.666	0.741	0.766	0.729	0.844	0.822	0.804	0.827	0.846	**0.888**	**0.935**
	E_ξ ↑	0.890	0.773	0.856	0.870	0.881	0.932	0.928	0.893	0.910	0.923	**0.943**	**0.973**
[8]	MAE ↓	0.208	0.111	0.090	0.093	0.068	0.055	0.065	0.049	0.046	0.038	**0.030**	**0.021**
LFSD	S_m ↑	0.736	0.753	0.698	0.791	0.738	0.796	0.787	0.794	0.801	0.828	**0.847**	**0.837**
	maxF ↑	0.726	0.817	0.791	0.817	0.744	0.791	0.771	0.779	0.796	0.826	**0.856**	**0.835**
	E_ξ ↑	0.804	0.856	0.840	0.865	0.815	0.865	0.839	0.835	0.847	0.872	**0.900**	**0.873**
[29]	MAE ↓	0.208	0.155	0.167	0.138	0.133	0.119	0.132	0.112	0.111	**0.088**	**0.075**	0.094
STERE	S_m ↑	0.660	0.731	0.708	0.757	0.825	0.848	0.873	0.875	0.871	0.879	**0.886**	**0.890**
	maxF ↑	0.633	0.740	0.755	0.757	0.823	0.831	0.863	0.860	0.861	0.874	**0.886**	**0.882**
	E_ξ ↑	0.787	0.819	0.846	0.847	0.887	0.912	0.927	0.925	0.923	0.925	**0.938**	**0.932**
[37]	MAE ↓	0.250	0.148	0.143	0.141	0.075	0.086	0.068	0.064	0.060	**0.051**	**0.047**	**0.051**
SSD	S_m ↑	0.621	0.704	0.675	0.747	0.714	0.776	0.813	0.841	0.839	0.807	**0.857**	**0.868**
	maxF ↑	0.619	0.711	0.710	0.735	0.687	0.729	0.781	0.807	0.810	0.766	**0.844**	**0.848**
	E_ξ ↑	0.736	0.786	0.800	0.828	0.807	0.865	0.882	0.894	0.897	0.852	**0.906**	**0.909**
[56]	MAE ↓	0.278	0.169	0.165	0.142	0.118	0.099	0.082	0.062	0.063	0.082	**0.058**	**0.052**
DUT-	S_m ↑	0.695	0.499	0.526	0.736	0.702	0.831	0.791	0.801	0.808	0.818	**0.889**	**0.903**
RGBD	maxF ↑	0.692	0.411	0.458	0.740	0.659	0.823	0.767	0.771	0.790	0.795	**0.898**	**0.901**
	E_ξ ↑	0.800	0.654	0.709	0.823	0.796	0.899	0.859	0.856	0.861	0.859	**0.933**	**0.937**
[40]	MAE ↓	0.220	0.243	0.201	0.144	0.122	0.097	0.113	0.100	0.093	0.076	**0.048**	**0.043**

i.e., NJUD, NLPR, RGBD135, and LFSD. The basic UNet model with the naive fusion decoder modules trained only on RGB images is used as the baseline model. The experimental results are shown in Table 1.

Effectiveness of the DenseASPP module. The first row in Table 1 denotes the baseline UNet, while the second row means we adopt the DenseASPP module. The comparison results show that using the DenseASPP module can moderately improve the saliency detection performance, offering us a more powerful baseline model for evaluating our proposed attention module and the residual fusion module.

Effectiveness of the S^2MA module. We show the model performance of further adding the NL module, the proposed SMA module, and the proposed S^2MA module in the 3^{rd} to 5^{th} rows in Table 1. We can see that adding the NL [51] module can only slightly improve (or even degrade) the model performance on the basis of a powerful baseline model (UNet+DenseASPP). Whereas using our proposed SMA module brings significant performance gain on the NJUD, RGBD135, and LFSD datasets, which demonstrates the effectiveness of the proposed self-mutual attention fusion scheme. Finally, using the proposed S^2MA module can further moderately improve the model performance, especially on the NLPR, RGBD135, and LFSD datasets. These results indicate that using the proposed selection attention to weight the mutual attention in attention fusion is beneficial. We also tried to use α^r and α^d to weight both self-attention and mutual attention but got worse results.

To thoroughly understand the effectiveness of our proposed attention fusion scheme, we show some visualization examples of the learned RGB-based self-attention, the depth-based self-attention, the selection attention α^d, and the fused attention A^r in Figure 4. We can see that usually the self-attention learned in each single modality is imperfect or even noisy, while A^r learned in the proposed S^2MA module can locate related positions of the query position more accurately by fusing the information from both modalities. For α^d, we find it tends to be small for pixels with large depth values. This is because large depth is usually coarse-grained, thus being less discriminative. On the contrary, α^d tends to be large for close pixels since their depth are more accurate and discriminative. Moreover, α^d are around 0.5 for pixels inside salient objects, which means RGB and depth attention are equally important for salient

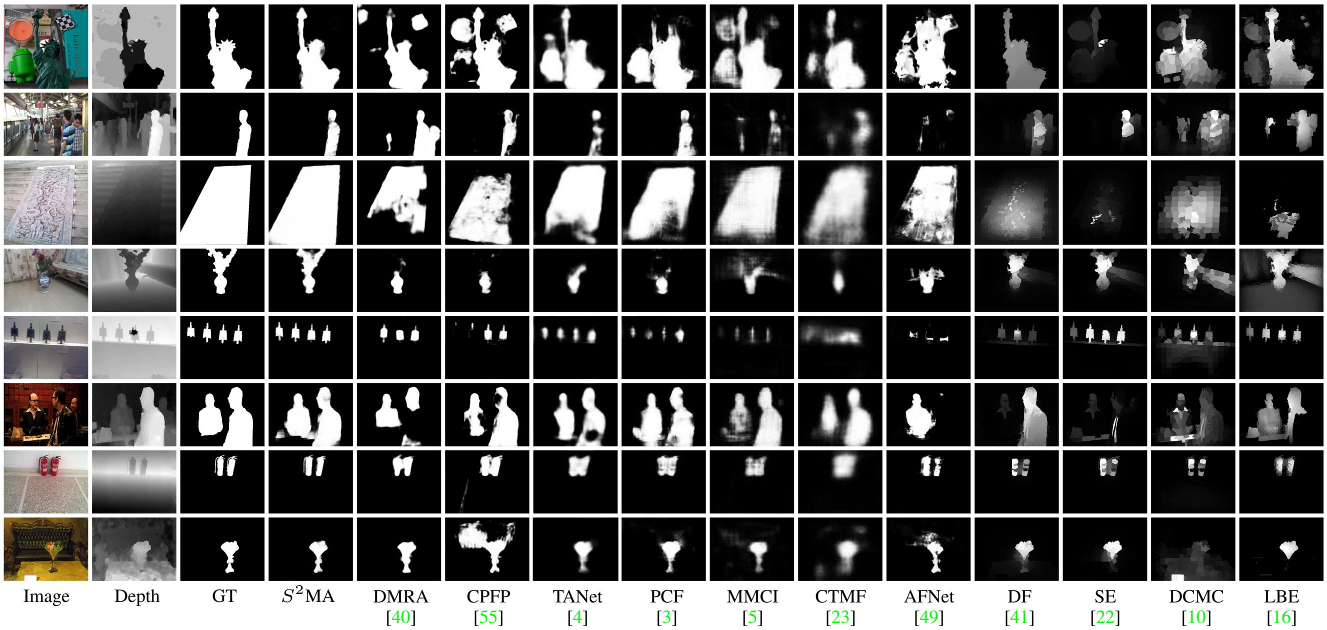


Figure 5. Qualitative comparison against 11 state-of-the-art RGB-D saliency detection methods. (GT: ground truth)

regions.

Effectiveness of the residual fusion module. In the last row of Table 1, we further adopt the residual fusion module Ⓡ in our saliency model. The results indicate that using this module to further fuse the depth decoder features into the RGB stream with a residual path can further improve the model performance, especially on the RGBD135 and the LFSD datasets.

5.5. Comparison with State-of-the-Art Methods

To evaluate the effectiveness of our proposed saliency model, we compare it with other 11 recently published RGB-D saliency methods, which include LBE [16], DCMC [10], SE [22], DF [41], AFNet [49], CTMF [23], MMCI [5], PCF [3], TANet [4], CPFP [55], and DMRA [40]. The first three methods are based on traditional models while the last eight ones are deep models. Since the training set of the DMRA [40] model further includes 800 images from the DUT-RGBD dataset, we further finetune our model on these images for a fair comparison on this dataset.

In Table 2 we show the quantitative comparison results. We can see that our model achieves the best performance on the NJUD, NLPR, RGBD135, SSD, and DUT-RGBD datasets. Especially on the RGBD135 dataset, the proposed S^2MA model outperforms the second-best model by a large margin. On the other two datasets, our model achieves the second-best performance but is close to the best model.

We also give qualitative comparison results in Figure 5. It shows that our model can handle various challenging scenarios, *e.g.*, the first two images have complex backgrounds, the salient objects in the 3^{rd} and the 4^{th} images have similar appearance with the backgrounds, the 5^{th} to the 7^{th} images have multiple objects. Generally, our model can accurately localize salient objects and segment them precisely, while other models are heavily disturbed in these complex scenes.

6. Conclusion

In this paper, we propose to fuse the self-attention and the other modality's attention in the Non-local model as a novel way for fusing multi-modal information. The fused attention is more accurate thus can propagate better global contexts. We also develop a selection attention mechanism to reweight the mutual attention term for filtering out unreliable modality information. The proposed S^2MA module is embedded into a two-stream CNN to solve the RGB-D saliency detection problem. Experimental results show that S^2MA significantly improves the model performance on the basis of a powerful baseline model. As a result, our saliency model performs favorably against state-of-the-art RGB-D saliency detection methods. In the future, our proposed S^2MA module can also be used for other multi-modal tasks, such as video saliency detection [15], visual questions and segmentation answers [20], and audio-visual tasks [21].

Acknowledgments

This work was supported in part by the National Key R&D Program under Grant 2017YFB1002201. This work was also supported by ”Research Funds for Interdisciplinary subject, NWPU” and the Fundamental Research Funds for the Central Universities(3102019PJ005).

References

[1] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv preprint arXiv:1904.11492*, 2019. 2

[2] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *BMVC*, 2017. 1

[3] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *CVPR*, pages 3051–3060, 2018. 2, 6, 7, 8

[4] Hao Chen and Youfu Li. Three-stream attention-aware network for rgb-d salient object detection. *TIP*, 28(6):2825–2835, 2019. 2, 7, 8

[5] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. *Pattern Recognition*, 86:376–385, 2019. 2, 7, 8

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

[7] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Aˆ 2-nets: Double attention networks. In *NIPS*, pages 352–361, 2018. 2

[8] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *International Conference on Internet Multimedia Computing and Service*, page 23. ACM, 2014. 2, 6, 7

[9] Arridhana Ciptadi, Tucker Hermans, and James Rehg. An in depth view of saliency. In *BMVC*, 2013. 2

[10] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, and Chunping Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *IEEE Signal Processing Letters*, 23(6):819–823, 2016. 2, 7, 8

[11] Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. Depth really matters: Improving visual salient region detection with depth. In *BMVC*, 2013. 1

[12] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4558–4567, 2017. 6

[13] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704. AAAI Press, 2018. 6

[14] Deng-Ping Fan, Zheng Lin, Jia-Xing Zhao, Yun Liu, Zhao Zhang, Qibin Hou, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, datasets, and large-scale benchmarks. *arXiv preprint arXiv:1907.06781*, 2019. 1, 2

[15] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019. 8

[16] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for rgb-d salient object detection. In *CVPR*, pages 2343–2350, 2016. 7, 8

[17] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *CVPR*, pages 1623–1632, 2019. 1

[18] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019. 2

[19] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *CVPR*, 2020. 2

[20] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *ICCV*, pages 1811–1820, 2017. 8

[21] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *ICCV*, pages 7053–7062, 2019. 8

[22] Jingfan Guo, Tongwei Ren, and Jia Bei. Salient object detection for rgb-d image via saliency evolution. In *ICME*, pages 1–6. IEEE, 2016. 1, 2, 7, 8

[23] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics*, 48(11):3171–3183, 2017. 2, 6, 7, 8

[24] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 2

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 4

[26] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 2

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 5

[28] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *ICIP*, pages 1115–1119. IEEE, 2014. 2, 5, 6, 7

[29] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *CVPR*, pages 2806–2813, 2014. 6, 7

[30] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 2019. 2

[31] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3917–3926, 2019. 1

[32] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, pages 678–686, 2016. 1

[33] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018. 1, 4, 6

[34] Zhengyi Liu, Song Shi, Quntao Duan, Wei Zhang, and Peng Zhao. Salient object detection for rgb-d image by single stream recurrent convolution neural network. *Neurocomputing*, 363:46–57, 2019. 1, 2

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 4

[36] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, pages 299–307, 2017. 3

[37] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461. IEEE, 2012. 6, 7

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 6

[39] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *ECCV*, pages 92–109. Springer, 2014. 6, 7

[40] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, 2019. 2, 6, 7, 8

[41] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *TIP*, 26(5):2274–2285, 2017. 1, 2, 7, 8

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 4, 5

[43] Carsten Rother, Lucas Bordeaux, Youssef Hamadi, and Andrew Blake. Autocollage. In *ACM Transactions on Graphics*, volume 25, pages 847–852, 2006. 1

[44] Riku Shigematsu, David Feng, Shaodi You, and Nick Barnes. Learning rgb-d salient object detection using background enclosure, depth contrast, and top-down features. In *ICCV Workshops*, pages 2749–2757, 2017. 2

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[46] Hangke Song, Zhi Liu, Huan Du, Guangling Sun, Olivier Le Meur, and Tongwei Ren. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *TIP*, 26(9):4204–4216, 2017. 1

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2

[48] Yao Wan, Jingdong Shu, Yulei Sui, Guandong Xu, Zhou Zhao, Jian Wu, and Philip Yu. Multi-modal attention network learning for semantic source code retrieval. In *IEEE/ACM International Conference on Automated Software Engineering*, pages 13–25. IEEE, 2019. 3

[49] Ningning Wang and Xiaojin Gong. Adaptive fusion for rgb-d salient object detection. *IEEE Access*, 7:55277–55284, 2019. 1, 2, 7, 8

[50] Wenguan Wang, Jianbing Shen, and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *TPAMI*, 41(7):1531–1544, 2018. 1

[51] Xiaolong Wang, Ross B Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 1, 2, 3, 4, 7

[52] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI*, 39(11):2314–2320, 2017. 1

[53] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, pages 3684–3692, 2018. 4

[54] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019. 2

[55] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *CVPR*, 2019. 2, 6, 7, 8

[56] Chunbiao Zhu and Ge Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *ICCV Workshops*, pages 3008–3014, 2017. 6, 7

[57] Zhen Zhu, Mengdu Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, 2019. 2