Deep RGB-D Saliency Detection without Depth

Yuan-fang Zhang, Jiangbin Zheng, Wenjing Jia, Wenfeng Huang,
Long Li, Nian Liu, Fei Li, and Xiangjian He, *Senior member, IEEE*

***Abstract*—The existing saliency detection models based on RGB colors only leverage appearance cues to detect salient objects. Depth information also plays a very important role in visual saliency detection and can supply complementary cues for saliency detection. Although many RGB-D saliency models have been proposed, they require to acquire depth data, which is expensive and not easy to get. In this paper, we propose to estimate depth information from monocular RGB images and leverage the intermediate depth features to enhance the saliency detection performance in a deep neural network framework. Specifically, we first use an encoder network to extract common features from each RGB image and then build two decoder networks for depth estimation and saliency detection, respectively. The depth decoder features can be fused with the RGB saliency features to enhance their capability. Furthermore, we also propose a novel dense multiscale fusion model to densely fuse multiscale depth and RGB features based on the dense ASPP model. A new global context branch is also added to boost the multiscale features. Experimental results demonstrate that the added depth cues and the proposed fusion model can both improve the saliency detection performance. Finally, our model not only outperforms state-of-the-art RGB saliency models, but also achieves comparable results compared with state-of-the-art RGB-D saliency models.**

***Index Terms*—Saliency detection, Depth estimation, Convolutional neural network, Feature fusion.**

I. Introduction

VISUAL saliency detection simulates the human visual attention mechanism to detect the most attractive objects in visual scenes. This task can be used as a pre-processing technique and help many other computer vision tasks to locate salient objects first, thus improving their effectiveness and efficiency.

Previous saliency models usually detect salient objects from RGB signals, which can be easily captured by modern cameras or cell phones. They usually use the contrast mechanism [1] to find the regions different from others and extract semantic features to find the regions which are most likely to be objects. Although recent CNN-based RGB saliency models [2], [3], [4], [5], [6] and other models [7], [8], [9], [10], [11] have achieved very promising results, they can still easily fail to detect salient objects in challenging scenarios since RGB data can only provide visual cues for saliency detection, thus greatly limiting the model capability. Figure 1 shows two examples. We can see that for images with cluttered backgrounds (the top row), saliency models that use RGB cues only (the column (c)) can be easily distracted by backgrounds. For salient objects with complex appearance (the bottom row), only using RGB cues (the column (c)) easily leads to incomplete detection.

On the other hand, the human visual system can easily leverage 3D information from the real world, in which the depth information plays a very important role in visual perception. For example, in the human visual attention mechanism, salient objects usually have different depths with the backgrounds. To this end, some researchers propose to detect saliency using RGB-D data. First, RGB-D images are captured using 3D sensors such as Microsoft Kinect, stereo cameras, light field cameras, etc. Then, they fuse RGB and depth saliency cues to obtain the saliency detection results. As such, depth maps can provide complementary information for appearance cues and thus promote the saliency detection performance, especially for challenging scenarios. Nevertheless, 3D sensors are not popular and usually expensive, making RGB-D images much more difficult to obtain than RGB images.

To solve this problem, in this paper we propose a novel deep learning framework to detect RGB-D saliency without actually requiring input depth data. Specifically, we predict depth maps for RGB images and simultaneously fuse depth features with RGB features to detect salient objects. By using the predicted depth information, our model can filter out the distraction from backgrounds (see the top row of Figure 1) and highlight the whole salient object more uniformly (see the bottom row of Figure 1). On the one hand, we leverage depth information to detect saliency more accurately. On the other hand, we do not require testing images to have depth maps and only use common RGB images. In Figure 1 (c) and (e), we show the comparison of the saliency detection results between our proposed model (column (e)) and a baseline model without using depth cues (column (c)). The results show that our model can obviously improve saliency detection performance for RGB images.

Furthermore, previous RGB-D saliency detection models usually fuse RGB features with depth features by using simple feature concatenation, addition [12], [13], or attention models

This work is sponsored by Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (CX201959) and Synergy Innovation Foundation of the University and Enterprise for Graduate Students in Northwestern Polytechnical University (XQ201910). This work is also supported in part by the National Natural Science Foundation of China under Grant 61972321. Jiangbin Zheng and Xiangjian He are the corresponding authors.

Yuan-fang Zhang is with School of Computer Science, Northwestern Polytechnical University, P.R.China and Faculty of Engineering and IT, University of Technology Sydney, Australia (Email: zyf.robinzhang@gmail.com).

Jiangbin Zheng is with School of Computer Science, Northwestern Polytechnical University, P.R.China (Email: zhengjb@nwpu.edu.cn).

Xiangjian He and Wenjing Jia are with Faculty of Engineering and IT, University of Technology Sydney, Australia (Email: Xiangjian.He@uts.edu.au, Wenjing.Jia@uts.edu.au).

Wenfeng Huang is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, P.R.China. Long Li is with School of Automation, Northwestern Polytechnical University, P.R.China. Nian Liu is with Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. Fei Li is with School of Software, Northwestern Polytechnical University, P.R.China. (email:huang-wenfeng@outlook.com, longli.nwpu@gmail.com, liunian228@gmail.com, foreverfei875@gmail.com)

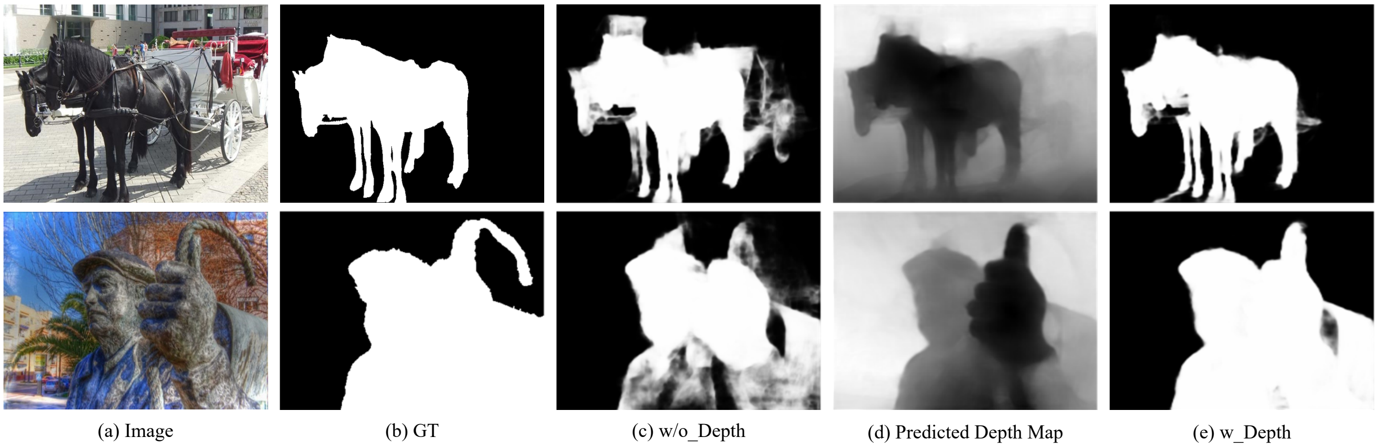


Fig. 1. **Comparison of the saliency detection results without ("w/o_Depth") and with ("w_Depth") using depth cues.** (a) and (b) show two example images and their ground truth (GT) saliency maps. (c) shows the saliency detection results of a baseline deep saliency model without using depth cues. (d) shows our predicted depth maps. (e) shows our predicted saliency maps with using depth cues.

[14], [15]. Different from them, inspired by the DenseASPP (DASPP) model [16], we propose a novel multimodal feature fusion model by densely fusing RGB and depth DASPP features, thus greatly enriching the feature fusion paths across multiple scales. Considering that the original DASPP model only incorporates multiscale local features, we also propose to enhance it with an additional global context propagation module [17]. Experimental results demonstrate that this proposed feature fusion model can better improve multimodal features. Finally, our saliency model outperforms previous RGB based saliency detection methods and also achieves comparable or even better results compared with state-of-the-art RGB-D saliency algorithms.

To sum up, our contributions of this work are as follows:

- We propose the first deep saliency model to leverage depth cues for enhancing the saliency detection performance but without actually using depth data.
- We propose a novel depth feature fusion model by introducing dense fusion paths in DASPP and also enhance it by incorporating global contexts.
- Experimental results demonstrate the effectiveness of our proposed model. It not only outperforms previous RGB saliency models, but also can obtain *comparable results with* state-of-the-art RGB-D saliency methods.

In the subsequent sections, we first discuss our model with related work in Section II. Then we present our proposed model in Section III in detail and report the experimental results in Section IV. Finally, we draw conclusion in Section V.

II. Related Work

A. RGB Saliency Detection

Early RGB saliency detection models usually extract low-level image features and then leverage the contrast mechanism [18], [19], background prior [20], [21], or objectness prior [22], [23] to detect salient objects. Recently, many research works introduced CNNs into the saliency detection field and have achieved very promising results. Most of these methods directly solve the saliency detection problem using end-to-end CNNs. For example, early models [24], [25], [26] usually use multi-scale CNNs to extract multi-scale features for each pixel or superpixel from its multiple local and global patches, and then combine the multi-scale deep features to classify or regress its saliency value. Subsequent models adopt the fully convolutional network (FCN) [27] architecture to perform saliency classification for each pixel simultaneously. Typically, encoder and decoder model has been a trend for researcher[28], they first use an encoder with pretrained parameters, such as VGG [29] or ResNet [30], to extract multi-level deep features, and then build a decoder to fuse these multi-level features for saliency detection. Some works [2], [31], [32], [33] use the U-Net [34] architecture to progressively fuse multi-level features, and some other works [35], [3] adopt the HED [36] network architecture to fuse them simultaneously. These above-mentioned methods all directly infer image saliency from extracted deep features, without considering other knowledge.

Some complementary knowledge have been introduced to enhance the saliency detection performance. Li *et al.* [37] introduced the semantic segmentation task to enhance the feature capability for object perception. Wang *et al.* [38] used eye fixation to guide the detection of salient objects. In [39], Zhang *et al.* leveraged image captioning to help to capture semantic information of salient objects in visual scenes. Recently, many deep saliency models [40], [41], [5], [42], [6] have proposed to simultaneously predict object contours and use the contour prior to enhance the object boundaries for salient objects. However, none of these works has explored using depth knowledge to enhance the saliency detection performance. In this work, we propose to simultaneously predict the depth map for each image and use the depth features to supplement the RGB features for saliency detection.

In [43], Xiao *et al.* also proposed to derive pseudo depth from the RGB images, and then leverage the pseudo depth to boost the performance of RGB saliency models by computing a depth-driven background prior and a depth contrast feature. Our method significantly differs from theirs in two aspects. First, their model is based on traditional saliency features and frameworks, while ours is an end-to-end deep saliency model, thus having much better model performance and much faster speed. Second, their model needs to derive the pseudo depth map first and then use it to compute the depth-based feature and prior map, while ours can use the intermediate depth features to boost the RGB features before the generation of the depth map, thus is more effective and efficient.

B. RGB-D Saliency Detection

Traditional RGB-D saliency models usually use the depth map as another channel and follow RGB saliency models to derive some saliency cues, such as depth-based contrast [44] or background priors. Finally, RGB and depth cues are combined to obtain the final saliency detection results. Some other models propose some special saliency cues from the depth data, such as the shape and 3D layout priors proposed in [45], to supplement the RGB saliency cues. Recently, many works adopted CNNs in the RGB-D saliency detection task and have obtained much better results than traditional models. Some of the models [46], [47] regard the depth map as the fourth channel besides the RGB image and then train a deep saliency model with four-channel input images. Some other models [48] adopt two CNNs on the RGB image and the depth map separately to generate two saliency maps and then fuse them to obtain the final saliency map. Most works use two-stream CNNs to extract RGB and depth features from the two modalities, respectively, and then fuse the multimodal features with various methods, such as feature concatenation and addition [12], [13], spatial channel attention [14], [15], and mutual attention [49]. All these methods require to input the captured depth maps into the saliency models for enhancing their performance. In contrast, we propose to infer depth maps from the input RGB images and simultaneously leverage the intermediate depth features to enhance the RGB features, thus exploiting the depth knowledge for saliency detection without actually requiring depth data. Furthermore, different from previous feature fusion schemes, we propose a novel feature fusion model by adding dense fusion paths and a global context propagation branch in DASPP.

A contemporary work in [50] also proposed to eliminate the dependency on depth maps for RGB-D saliency detection. They trained a saliency detection branch based on depth data, and then used the result to perform knowledge distillation for promoting the model capability of the RGB saliency branch. Although our model and theirs try to achieve the same goal, the implementation mechanism is totally different. First, they aim at designing a light-weight saliency detection model and adopt the knowledge distillation technique, while we aim to build a powerful saliency model and also eliminate the dependency on the input depth maps, hence performing multi-task learning. As a result, they can only leverage RGB-D saliency detection data to train their model, while ours can exploit large-scale external RGB saliency detection data and depth estimation data. Second, their model is implemented based on knowledge distillation, hence the model capacity is theoretically limited by the teacher network, i.e., the depth saliency detection branch. In contrast, our method explicitly fuses RGB information with the inferred depth feature. Thus, the model capacity is an ensemble of both modalities. As a result, our model is much more effective than theirs, although theirs maybe more efficient.

III. Proposed Method

In this part, we articulate the proposed deep saliency model in detail. As shown in Figure 2, given each image, we first use an encoder network to extract multi-level encoder features. Then we follow the U-Net [34] architecture to progressively fuse the multi-level features to predict the depth map and the saliency map via two decoder networks, respectively. We also fuse the depth features with the RGB features to leverage depth cues for enhancing the saliency detection performance. Specifically, we adopt a DASPP [16] model at the beginning of the depth decoder branch, and then fuse the depth DASPP features with RGB features in a novel Dense MultiScale Fusion (DMSF) module. Subsequently, we fuse each depth decoding feature map with each corresponding RGB decoding feature map.

A. Encoder Network

Following [4], our encoder network is based on the VGG-16 network [29]. It is an FCN and has seven convolutional (Conv) blocks. The first five blocks are based on the five Conv blocks of VGG-16, i.e., Conv1-Conv5, each of which is composed of two or three consecutive Conv layers. The last two blocks are based on the two fully connected (FC) layers of VGG-16, i.e., FC6 and FC7. Since the original VGG-16 network has five pooling layers following the five Conv blocks, respectively, thus downsampling the input image by a factor of 32, which is too large for saliency detection. To enlarge the spatial sizes of the feature maps, we change the strides of the last two pooling layers to 1, and also use atrous Conv layers [51] with rate $r = 2$ in the Conv5 block. We also transform the two FC layers of VGG-16 to Conv layers for taking advantage of the plentiful features learned in them. Concretely, the FC6 layer is transformed to a 3×3 atrous Conv layer with $r = 12$ and 1024 channels, while the FC7 layer is transformed to a 1×1 Conv layer with the same channel number. Finally, we obtain the final FC7 feature map with a downsampling factor of 8, and also get five multi-level feature maps from Conv1-Conv5.

B. Decoder Networks

Next, we construct two decoder branches for predicting the depth map and the saliency map. We name them as the depth branch and the RGB saliency branch, respectively. We first use DASPP and DMSF to extract and fuse multiscale features from the FC7 encoder features, and then follow the U-Net [34] architecture to progressively fuse multi-level encoding features in subsequent decoding modules.

1) DASPP and DMSF: For each decoder branch, we first use a Conv layer to reduce the channel number of the FC7 feature map to 512 channels. We denote the two feature maps as $\boldsymbol{X}^R_{FC7'}$ for the RGB saliency branch and $\boldsymbol{X}^D_{FC7'}$ for the depth branch. Then we extract and fuse multiscale features based on the DASPP [16] model. Specifically, we

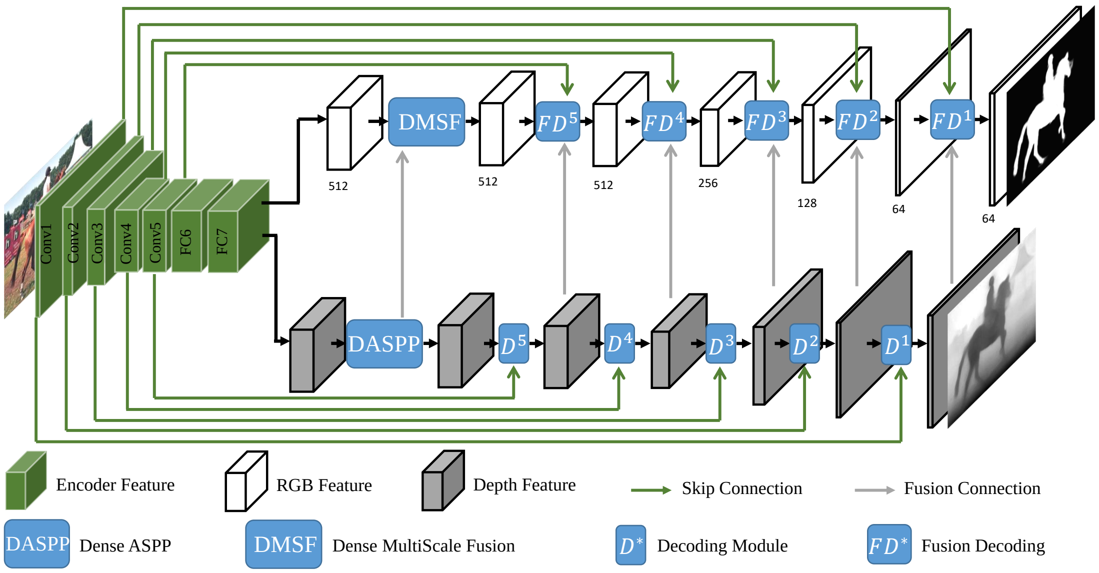


Fig. 2. **Network architecture of the proposed model.** The whole model has an encoder network (green cubes) and two decoder networks (white and gray cubes). The encoder network is used to extract multi-level encoder features, while the two decoder networks are used for predicting the depth map and the saliency map, respectively. We use the VGG-16 network [29] as our encoder, and its multi-level features are marked on the cubes. Each decoder progressively fuses the multi-level features by using skip-connections. The depth features are also fused with the RGB features via fusion connections for enhancing the saliency detection performance. The channel numbers of the decoding modules are also marked under the cubes.

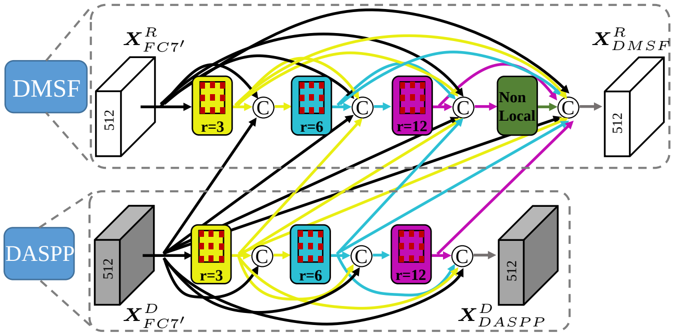


Fig. 3. **Network architecture of the DASPP model and the proposed DMSF model.**

directly use DASPP for the depth feature $\boldsymbol{X}^D_{FC7'}$. DASPP deploys several atrous Conv layers on the input feature map with increasing rates, thus obtaining multiscale features with different receptive fields. Meanwhile, it also introduces dense connections among the multiscale atrous layers, connecting each layer to all subsequent layers with larger rates, thus covering scale ranges densely.

As shown in Figure 3, we use three atrous layers with rates $r = \{3, 6, 12\}$. In each layer $i \in \{0, 1, 2\}$, we first concatenate all previous features. Then, a 1×1 Conv layer is used to reduce the channel number to 512. Finally, we use an atrous Conv layer with rate r_i to generate the atrous feature $\boldsymbol{X}^D_i$ with a large receptive field. We follow [16] to set the channel number of each atrous Conv layer to $\lfloor \frac{512}{3} \rfloor$ for reducing the computational cost. The whole process can be written as:

$$\boldsymbol{X}^D_i = \mathbb{AC}_{r_i}(\mathbb{C}([\boldsymbol{X}^D_{FC7'}|\boldsymbol{X}^D_0|\cdots|\boldsymbol{X}^D_{i-1}])), \tag{1}$$

where $\mathbb{C}$ means a Conv layer while $\mathbb{AC}_{r_i}$ means an atrous Conv layer with rate r_i. $[\cdot|\cdot]$ denotes the concatenation operation.

Finally, the three multiscale features and the original feature are concatenated to form the final depth DASPP feature via a 1×1 Conv layer with 512 channels as:

$$\boldsymbol{X}^D_{DASPP} = \mathbb{C}([\boldsymbol{X}^D_{FC7'}|\boldsymbol{X}^D_0|\boldsymbol{X}^D_1|\boldsymbol{X}^D_2]). \tag{2}$$

Nevertheless, DASPP is only designed for a single modality. To adapt it to multimodal features, we propose a novel dense multiscale fusion (DMSF) model by extending DASPP with dense fusion connections to fuse cross modality features. In our case, we densely fuse the depth multiscale features with the RGB ones. Specifically, we deploy the same three atrous Conv layers on the input RGB feature map. At the same time, we not only densely connect each RGB atrous feature $\boldsymbol{X}^R_i$ to all subsequent atrous layers, but also densely connect each depth atrous feature $\boldsymbol{X}^D_i$ to all RGB atrous layers with larger rates as:

$$\begin{aligned}\boldsymbol{X}^R_i = \mathbb{AC}_{r_i}(\mathbb{C}([&\boldsymbol{X}^R_{FC7'}|\boldsymbol{X}^R_0|\cdots|\boldsymbol{X}^R_{i-1}| \\ &\boldsymbol{X}^D_{FC7'}|\boldsymbol{X}^D_0|\cdots|\boldsymbol{X}^D_{i-1}])).\end{aligned} \tag{3}$$

Furthermore, since DASPP only uses atrous Conv layers to construct multiscale features, these features all have local contexts (since the atrous Conv operation is a local operation). Based on the basic idea of the DASPP model to construct multiscale features with small to large scales, we propose to incorporate global contexts at the end of DMSF as the largest scale. Specifically, we adopt the non-Local network [17] as the global context model since its effectiveness has been widely verified. For the input feature map $\boldsymbol{X} \in \mathbb{R}^{W \times H \times C}$, the non-local network computes three feature embeddings $\theta(\boldsymbol{X}), \phi(\boldsymbol{X}), g(\boldsymbol{X}) \in \mathbb{R}^{WH \times C'}$ first. Then it uses $\theta(\boldsymbol{X})$ and $\phi(\boldsymbol{X})$ to compute an global attention matrix with size $WH \times WH$, which can be used to propagate global contexts from $g(\boldsymbol{X})$. Finally, the global contexts are transformed to C'' channels by a Conv layer.

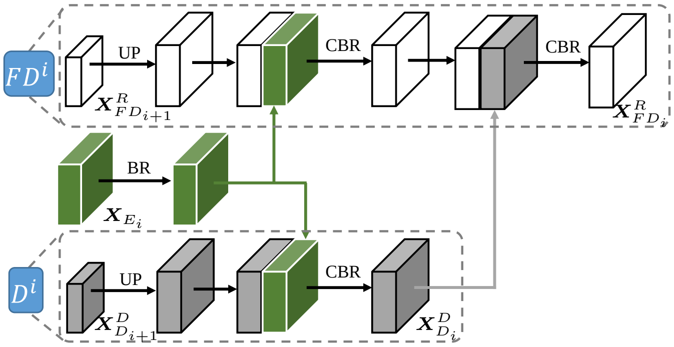

FD^i
UP
CBR
CBR
$\boldsymbol{X}^R_{FD_{i+1}}$
$\boldsymbol{X}^R_{FD_i}$
BR
$\boldsymbol{X}_{E_i}$
D^i
UP
CBR
$\boldsymbol{X}^D_{D_{i+1}}$
$\boldsymbol{X}^D_{D_i}$

Fig. 4. **Network architecture of the decoding module and the fusion decoding module.** "BR" means BN [52] and ReLU, "CBR" means Conv, BN and ReLU. "UP" means upsampling.

The whole model can be formulated as:

$$\mathbb{NL}(\boldsymbol{X}) = \mathbb{C}(softmax(\theta(\boldsymbol{X})\phi(\boldsymbol{X})^\top)g(\boldsymbol{X})), \tag{4}$$

where $softmax$ operates on each row of the attention matrix.

As shown in Figure 3, we adopt the non-local network as the forth multiscale layer in DMSF. It takes all previous RGB multiscale features and depth ones as inputs and propagates global contexts for them. We set the channel C' as 512 and C'' as $\lfloor \frac{512}{3} \rfloor$ to keep consistent with previous layers. The obtained feature is:

$$\boldsymbol{X}^R_{NL} = \mathbb{NL}([\boldsymbol{X}^R_{FC7'}|\boldsymbol{X}^R_0|\boldsymbol{X}^R_1|\boldsymbol{X}^R_2|\boldsymbol{X}^D_{FC7'}|\boldsymbol{X}^D_0|\boldsymbol{X}^D_1|\boldsymbol{X}^D_2]). \tag{5}$$

Finally, similar to DASPP, all previous RGB and depth layers are concatenated and a 1×1 Conv layer is used to obtain the final DMSF feature with 512 channels as:

$$\begin{aligned}\boldsymbol{X}^R_{DMSF} = \mathbb{C}([&\boldsymbol{X}^R_{FC7'}|\boldsymbol{X}^R_0|\boldsymbol{X}^R_1|\boldsymbol{X}^R_2|\boldsymbol{X}^R_{NL}| \\ &\boldsymbol{X}^D_{FC7'}|\boldsymbol{X}^D_0|\boldsymbol{X}^D_1|\boldsymbol{X}^D_2]).\end{aligned} \tag{6}$$

The final features $\boldsymbol{X}^D_{DASPP}$ $\boldsymbol{X}^R_{DMSF}$ extract and fuse multiscale features from the two FC7 features, thus supplying good starting points for the depth prediction branch and the saliency detection branch.

2) Decoding Modules: Inspired by U-Net, we design five decoding modules for each decoder branch to progressively fuse multi-level encoding features via skip-connections. At the same time, we fuse each depth decoding feature with the corresponding RGB one to enhance it for saliency detection. We name the decoding modules in the depth branch as D^i, where i is in inverse order from 5 to 1, as shown in Figure 2. For the RGB branch, we name them as fusion decoding modules FD^i since they simultaneously fuse encoding features and depth decoding ones. The encoding feature of the i^{th} Conv block is denoted as $\boldsymbol{X}_{E_i}$, which is the last Conv feature before using the ReLU activation function.

As shown in Figure 4, in the depth branch, each decoding module D^i fuses $\boldsymbol{X}_{E_i}$ with the previous depth decoding feature $\boldsymbol{X}^D_{D_{i+1}}$ to obtain the current depth decoding feature $\boldsymbol{X}^D_{D_i}$:

$$\boldsymbol{X}^D_{D_i} = \mathbb{CBR}([\mathbb{BR}(\boldsymbol{X}_{E_i})|\mathbb{UP}(\boldsymbol{X}^D_{D_{i+1}})]), \tag{7}$$

where $\mathbb{BR}$ means batch normalization (BN) [52] and ReLU. Here BN is used to normalize $\boldsymbol{X}_{E_i}$ for making it compatible with $\boldsymbol{X}^D_{D_{i+1}}$. $\mathbb{UP}$ denotes upsampling since $\boldsymbol{X}^D_{D_{i+1}}$ is smaller than $\boldsymbol{X}_{E_i}$ for $i \leq 3$. $\mathbb{CBR}$ means 1×1 Conv, BN, and ReLU. The channel numbers of $\boldsymbol{X}^D_{D_i}$ are set as 512, 256, 128, 64, 64 for $i \in \{5, 4, 3, 2, 1\}$, respectively. We use $\boldsymbol{X}^D_{DASPP}$ as the first decoding feature $\boldsymbol{X}^D_{D_6}$.

In the RGB saliency branch, each fusion decoding module FD^i fuses $\boldsymbol{X}_{E_i}$ with the previous RGB decoding feature $\boldsymbol{X}^R_{FD_{i+1}}$ to obtain the current RGB decoding feature $\boldsymbol{X}^R_{D_i}$. Then $\boldsymbol{X}^D_{D_i}$ is also fused to obtain $\boldsymbol{X}^R_{FD_i}$:

$$\begin{aligned}\boldsymbol{X}^R_{D_i} &= \mathbb{CBR}([\mathbb{BR}(\boldsymbol{X}_{E_i})|\mathbb{UP}(\boldsymbol{X}^R_{FD_{i+1}})]), \\ \boldsymbol{X}^R_{FD_i} &= \mathbb{CBR}([\boldsymbol{X}^R_{D_i}|\boldsymbol{X}^D_{D_i}]),\end{aligned} \tag{8}$$

where $\boldsymbol{X}^R_{DMSF}$ is used as $\boldsymbol{X}^R_{FD_6}$.

Finally, we directly use a 1×1 Conv layer with 1 channel on $\boldsymbol{X}^D_{D_1}$ to obtain the predicted depth map. A same Conv layer with the Sigmoid activation function is also used on $\boldsymbol{X}^R_{FD_1}$ to predict the saliency map.

C. Loss Functions

For saliency prediction, we use a simple binary cross entropy loss function. Suppose we have a predicted saliency map $\bar{S} \in [0, 1]^{W \times H}$ and the corresponding ground truth $S \in \{0, 1\}^{W \times H}$, then the saliency loss can be computed as:

$$L_s(\bar{S}, S) = \frac{1}{WH} \sum_{w,h=1}^{W,H} (S_{wh} log \bar{S}_{wh} + (1 - S_{wh}) log(1 - \bar{S}_{wh})). \tag{9}$$

To ease the network training, we also predict a saliency map from $\boldsymbol{X}^R_{FD_i}$ and adopt the supervision with the saliency loss for each fusion decoding module.

For depth prediction, we adopt the depth ranking loss in [53], which optimizes the ordinal relation of each pair of pixels. First, for each predicted depth map Z and the corresponding ground truth depth map G, we sample N point pairs. For pair k, we denote the pair of point as (i_k, j_k), where i_k and j_k are the coordinates of the two points. Its ordinal relation label ℓ_k can be defined as:

$$\ell_k = \begin{cases} +1, & \frac{G_{ik}}{G_{jk}} > 1 + \delta, \\ -1, & \frac{G_{jk}}{G_{ik}} > 1 + \delta, \\ 0, & otherwise, \end{cases} \tag{10}$$

where δ is an empirical threshold.

We follow [53] to set $N = 3000$ and $\delta = 0.02$. Then the depth ranking loss is defined as:

$$L_r(Z, G) = \frac{1}{\sum \omega_k} \sum_{k=1}^{N} \omega_k \psi(i_k, j_k, \ell_k, Z), \tag{11}$$

where

$$\psi = \begin{cases} log(1 + exp[(-Z_{ik} + Z_{jk})\ell_k]), & \ell_k \neq 0 \\ (Z_{ik} - Z_{jk})^2, & \ell_k = 0. \end{cases} \tag{12}$$

$\omega_k \in \{0, 1\}$ is the loss weight for pair k. We follow [53] to sort the losses ψ for all training pairs at each iteration, and set the pairs with the smallest 25% losses to have $\omega_k = 0$. In this way, we can increase the ratio of equal pairs and avoid keeping optimizing pairs with large difference.

However, using the ranking loss will lead to slow convergence. Thus we also adopt a normalized ℓ_2 loss between Z and G. Specifically, it is the ℓ_2 loss between the normalized Z and normalized G:

$$L_n(Z,G) = \frac{1}{WH}\sum_{w,h=1}^{W,H}(\frac{Z_{wh}-\mu_Z}{\sqrt{\sigma_Z}} - \frac{G_{wh}-\mu_G}{\sqrt{\sigma_G}})^2, \quad (13)$$

where μ_* and σ_* are the mean and variance, respectively.

For each decoding module in the depth decoder, we predict a depth map from $\boldsymbol{X}_{D_i}^D$ and use this loss to accelerate the network training.

IV. Experiments

In this section, we report the experimental results on *seven* RGB-D saliency benchmark datasets to validate the effectiveness of our proposed model.

A. Datasets and Evaluation Metrics

We conduct experiments *on widely used seven* RGB-D saliency detection datasets and four evaluation metrics.

The first dataset is the **NJUD** [54] dataset with 1985 stereo images. The second and the third one are the **NLPR** [55] and the **RGBD135** [44] dataset with 1000 and 135 RGB-D image pairs, respectively. These two datasets are both collected by using Microsoft Kinect. The forth dataset is the **SSD** [56] dataset with 80 stereo images collected from movies. *The fifth dataset is **DUT-RGBD**, which contains 800 training images and 400 testing images with real life scenarios. The sixth dataset is **STEREO** with 1,000 stereo images collected from the Internet. The last one is the **LFSD** dataset. It has 100 images captured by a light field camera.*

As for the evaluation metrics, the first one is the maximum F-measure (maxF) score. By binarizing the predicted saliency map with a threshold, we can compare it with the corresponding ground truth saliency map and obtain precision and recall. Then the F-measure score can be computed as:

$$F_m = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}, \quad (14)$$

where we follow previous work and set $\beta^2 = 0.3$ for emphasizing more on precision. By varying the threshold, we can find the maximum F-measure score. The second metric we used is the Mean Absolute Error (MAE). It computes the absolute difference between the predicted saliency map $\bar{S}$ and the ground truth S:

$$MAE = \frac{1}{WH}\sum_{w,h=1}^{W,H}|\bar{S}_{wh} - S_{wh}|. \quad (15)$$

The above two metrics both evaluate each pixel separately, without considering high-level statistics. Thus we also adopt the structure-measure S_m [57] score as the third metric. It computes and combines a region-aware structural similarity S_r and an object-aware one S_o between each saliency map and the ground truth as:

$$S_m = \alpha * S_o + (1-\alpha) * S_r, \quad (16)$$

where α is set to 0.5 by following the advice from [57].

The last metric is the recently proposed enhanced-alignment measure E_ξ [58]. It considers both global statistics and local pixel matching information. We use it for a more comprehensive evaluation. And our designed algorithm performance is displayed in Table I.

B. Implementation Details

We train our model in two stages using the stochastic gradient descent (SGD) algorithm. In the first stage, we leverage a depth estimation dataset, i.e., ReDWeb [53], and a saliency detection dataset, i.e., DUTS [59], to pretrain the model iteratively. The ReDWeb dataset contains 3600 stereo images collected from the Internet with various scenes, such as street, office, hill, park, etc. The DUTS dataset consists of 10553 training images collected from the ImageNet DET training/val sets [60] with human-annotated saliency maps. We first initialize the encoder part using the VGG-16 parameters pretrained on Imagenet and randomly initialize the two decoder branches. Then, for each iteration, we first use the ReDWeb data to train the encoder and the depth decoder branch with the depth losses L_r and L_n, then we use the DUTS data to train the whole network using the saliency loss L_s. We set the batch size and momentum to 10 and 0.9, respectively. The learning rates of the encoder part and the two decoders are set to 0.001 and 0.01, respectively. The whole training step is set to 40000 and we divide the learning rates by 10 at the 20000^{th} and 30000^{th} step.

In the second stage, we follow most previous works [12], [13], [15] to train the whole network using 1400 images of the NJUD dataset and 650 images of the NLPR dataset with using L_s and L_r losses. Since many depth maps of the NJUD dataset are very noisy, we do not use deep supervision with the L_n losses for the depth branch. We initialize the network parameters from the pretrained model in the first stage. The learning rates of the encoder part and the depth decoder are set to 0.0001 and the learning rate of the saliency decoder is set to 0.001 to fine-tune the network. Other training settings are set to the same as in the first stage.

We use the scale of 224×224 to train and test the network. Specifically, when training we first resize each RGB-D image pair to a random size from 224×224 to 272×272 and then randomly crop a 224×224 patch from it for network training. Random horizontal-flipping is also used for data augmentation. When testing, we directly resize each RGB-D image pair to 224×224 as the network input and then obtain the saliency map from the last layer of the RGB saliency branch. Each image is also pre-processed by subtracting the mean pixel value. The whole model is implemented using Pytorch [61]. *A GTX 1080Ti GPU is used for acceleration and the inference time for each testing image is only 0.019 second.*

C. Comparison with State-of-the-art Models

We compare our method with nine state-of-the-art saliency detection models, i.e., Amulet [32], DSS [3], BMP [62], PiCANet [4], R3Net [63], CPD [64], EGNet [6], *MINet [65], and ITSD [66].* We also include 12 state-of-the-art RGB-D

TABLE I

Quantitative comparison between our proposed model and state-of-the-art RGB and RGB-D salient object detection models. We compare our model with nine state-of-the-art (SOTA) CNN based RGB saliency models and *twelve SOTA deep learning based RGB-D saliency models on seven datasets in terms of four evaluation metrics.* “Train w D” means training with depth while “Test w D” means test with depth. Blue indicates the best performance in each group (i.e., RGB and RGB-D). Red indicates the cases our model outperforms RGB SOTA models, *while underline indicates the cases our model outperforms the A2dele model. “-” means the results are unavailable since the authors did not release them.*

	Train w D	Test w D	NJUD [54]				NLPR [55]				SSD [56]				RGBD135 [44]			
			S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
RGB Saliency Detection Models																		
Amulet [32]	✗	✗	0.827	0.819	0.879	0.079	0.838	0.779	0.885	0.055	0.822	0.808	0.876	0.077	0.823	0.761	0.872	0.065
DSS [3]	✗	✗	0.767	0.762	0.83	0.116	0.832	0.797	0.892	0.057	0.634	0.589	0.729	0.167	0.752	0.741	0.871	0.076
BMP [62]	✗	✗	0.860	0.850	0.905	0.068	0.880	0.848	0.917	0.045	0.849	0.817	0.895	0.060	0.878	0.854	0.920	0.040
PiCANet [4]	✗	✗	0.872	0.860	0.910	0.068	0.871	0.830	0.900	0.054	0.846	0.810	0.889	0.069	0.890	0.866	0.922	0.039
R3Net [63]	✗	✗	0.770	0.752	0.827	0.116	0.846	0.812	0.899	0.056	0.679	0.656	0.773	0.148	0.855	0.814	0.911	0.052
CPD [64]	✗	✗	0.863	0.858	0.905	0.060	0.893	0.866	0.925	0.034	0.833	0.804	0.878	0.067	0.896	0.882	0.932	0.028
EGNet [6]	✗	✗	0.840	0.826	0.883	0.079	0.880	0.847	0.917	0.045	0.740	0.701	0.802	0.126	0.888	0.872	0.919	0.036
MINet [65]	✗	✗	0.870	0.859	0.906	0.057	0.886	0.854	0.914	0.041	0.856	0.827	0.902	0.054	0.894	0.880	0.924	0.029
ITSD [66]	✗	✗	0.873	0.867	0.911	0.057	0.884	0.850	0.919	0.039	0.850	0.829	0.904	0.057	0.896	0.879	0.930	0.031
A2dele [50]	✓	✗	0.871	0.874	0.916	0.051	0.898	0.882	0.944	0.029	0.802	0.776	0.862	0.070	0.886	0.872	0.920	0.029
Ours	✓	✗	0.886	0.876	0.927	0.050	0.906	0.882	0.936	0.038	0.861	0.832	0.917	0.049	0.906	0.886	0.943	0.027
RGB-D Saliency Detection Models																		
DF [46]	✓	✓	0.763	0.804	0.864	0.141	0.802	0.778	0.880	0.085	0.747	0.735	0.828	0.142	0.752	0.766	0.870	0.093
AFNet [48]	✓	✓	0.772	0.775	0.853	0.100	0.799	0.771	0.879	0.058	0.714	0.687	0.807	0.118	0.770	0.729	0.881	0.068
CTMF [12]	✓	✓	0.849	0.845	0.913	0.085	0.860	0.825	0.929	0.056	0.776	0.729	0.865	0.099	0.863	0.844	0.932	0.055
MMCI [67]	✓	✓	0.858	0.852	0.915	0.079	0.856	0.815	0.913	0.059	0.813	0.781	0.882	0.082	0.848	0.822	0.928	0.065
PCF [13]	✓	✓	0.877	0.872	0.924	0.059	0.874	0.841	0.925	0.044	0.841	0.807	0.894	0.062	0.842	0.804	0.893	0.049
TANet [68]	✓	✓	0.878	0.874	0.925	0.060	0.886	0.863	0.941	0.041	0.839	0.810	0.897	0.063	0.858	0.827	0.910	0.046
CPFP [15]	✓	✓	0.878	0.877	0.923	0.053	0.888	0.867	0.932	0.036	0.807	0.766	0.852	0.082	0.872	0.846	0.923	0.038
DMRA [14]	✓	✓	0.886	0.886	0.927	0.051	0.899	0.879	0.947	0.031	0.857	0.844	0.906	0.058	0.900	0.888	0.943	0.030
SSF [69]	✓	✓	0.899	0.896	0.935	0.043	0.914	0.896	0.953	0.026	0.790	0.762	0.867	0.084	0.904	0.884	0.941	0.026
UCNet [70]	✓	✓	0.897	0.895	0.936	0.043	0.920	0.903	0.956	0.025	0.865	0.855	0.907	0.049	0.933	0.930	0.976	0.018
JLDCF [71]	✓	✓	0.897	0.899	0.939	0.044	0.920	0.907	0.959	0.026	-	-	-	-	0.913	0.905	0.955	0.026

	Train w D	Test w D	DUT-RGBD [14]				STEREO [72]				LFSD [73]			
			S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
RGB Saliency Detection Models														
Amulet [32]	✗	✗	0.813	0.792	0.875	0.089	0.867	0.854	0.919	0.053	0.804	0.808	0.865	0.100
DSS [3]	✗	✗	0.803	0.776	0.850	0.097	0.794	0.791	0.866	0.094	0.791	0.784	0.837	0.116
BMP [62]	✗	✗	0.855	0.843	0.890	0.069	0.891	0.880	0.931	0.049	0.802	0.790	0.844	0.103
PiCANet [4]	✗	✗	0.878	0.868	0.910	0.070	0.896	0.884	0.932	0.051	0.824	0.810	0.854	0.106
R3Net [63]	✗	✗	0.819	0.805	0.868	0.113	0.768	0.757	0.831	0.107	0.828	0.818	0.871	0.098
CPD [64]	✗	✗	0.875	0.865	0.911	0.055	0.893	0.886	0.929	0.042	0.822	0.811	0.860	0.089
EGNet [6]	✗	✗	0.872	0.853	0.905	0.059	0.859	0.844	0.903	0.063	0.834	0.829	0.869	0.090
MINet [65]	✗	✗	0.875	0.861	0.900	0.058	0.820	0.842	0.896	0.070	0.813	0.791	0.841	0.096
ITSD [66]	✗	✗	0.881	0.873	0.918	0.055	0.894	0.887	0.930	0.045	0.811	0.797	0.850	0.095
A2dele [50]	✓	✗	0.885	0.892	0.930	0.042	0.879	0.879	0.928	0.044	0.833	0.832	0.874	0.077
Ours	✓	✗	0.864	0.853	0.902	0.072	0.899	0.887	0.933	0.046	0.827	0.813	0.866	0.092
RGB-D Saliency Detection Models														
DF [46]	✓	✓	0.736	0.740	0.823	0.144	0.757	0.757	0.847	0.141	0.791	0.817	0.865	0.138
AFNet [48]	✓	✓	0.702	0.659	0.796	0.122	0.825	0.823	0.887	0.075	0.738	0.744	0.815	0.133
CTMF [12]	✓	✓	0.831	0.823	0.899	0.097	0.848	0.831	0.912	0.086	0.796	0.791	0.865	0.119
MMCI [67]	✓	✓	0.791	0.767	0.859	0.113	0.873	0.863	0.927	0.068	0.787	0.771	0.839	0.132
PCF [13]	✓	✓	0.801	0.771	0.856	0.100	0.875	0.860	0.925	0.064	0.794	0.779	0.835	0.112
TANet [68]	✓	✓	0.808	0.790	0.861	0.093	0.871	0.861	0.923	0.060	0.801	0.796	0.847	0.111
CPFP [15]	✓	✓	0.818	0.795	0.859	0.076	0.879	0.874	0.925	0.051	0.828	0.826	0.872	0.088
DMRA [14]	✓	✓	0.889	0.898	0.933	0.048	0.834	0.847	0.910	0.066	0.847	0.856	0.900	0.075
SSF [69]	✓	✓	0.915	0.924	0.951	0.033	0.837	0.840	0.912	0.065	0.859	0.867	0.900	0.066
UCNet [70]	✓	✓	0.871	0.866	0.910	0.059	0.903	0.899	0.944	0.039	0.864	0.864	0.905	0.066
JLDCF [71]	✓	✓	-	-	-	-	0.894	0.889	0.938	0.046	0.833	0.840	0.877	0.091

saliency detection models for comparison, including DF [46], AFNet [48], CTMF [12], MMCI [67], PCF [13], TANet [68], CPFP [15], DMRA [14], *SSF [69], UCNet [70], JLDCF [71], and A2dele [50]. Please note that all these models are deep learning based models and the last four are published in 2020. Since our model uses the VGG-16 network as the backbone,*

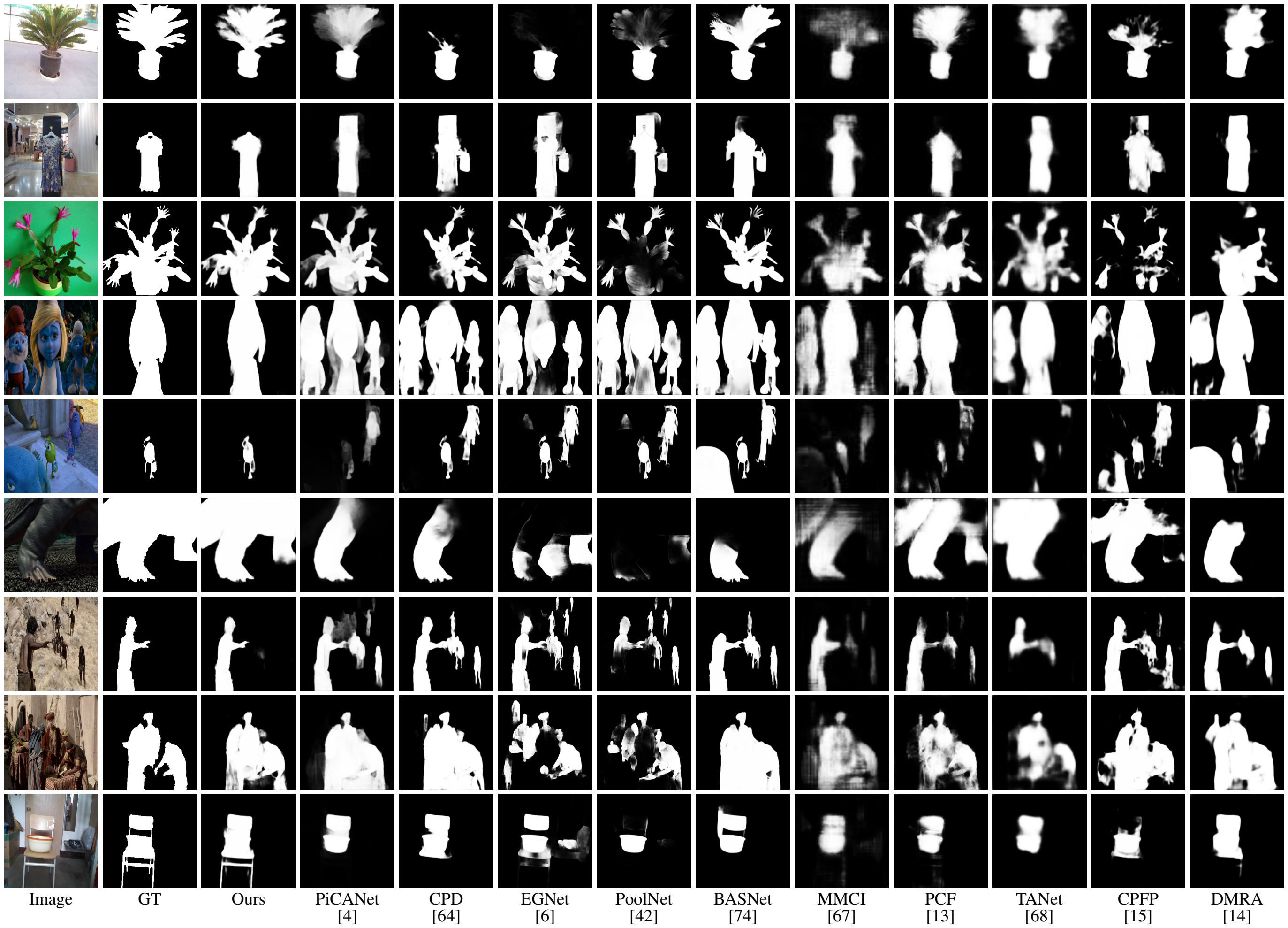


Fig. 5. **Visual comparison between our model and state-of-the-art RGB and RGB-D saliency models.** Our model outperforms SOTA RGB saliency models and surprisingly achieve comparable or even better results than SOTA RGB-D saliency models.

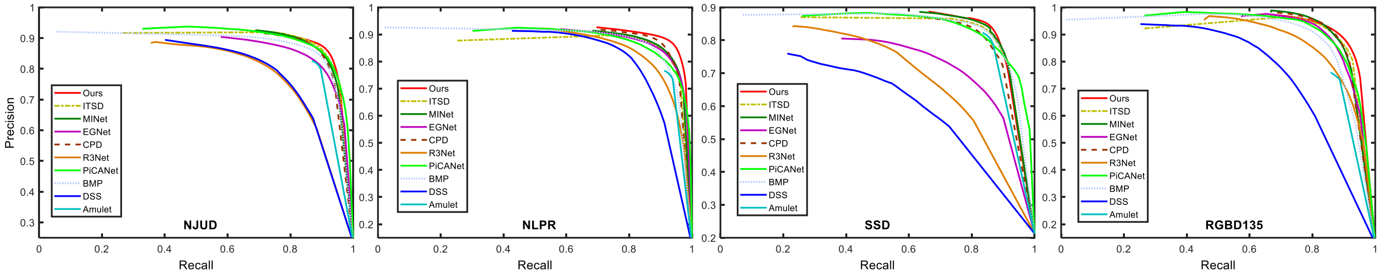


Fig. 6. **Comparison with state-of-the-art RGB saliency models in terms of PR curves on four datasets.**

we use the results of these models with the same backbone for fair comparisons.

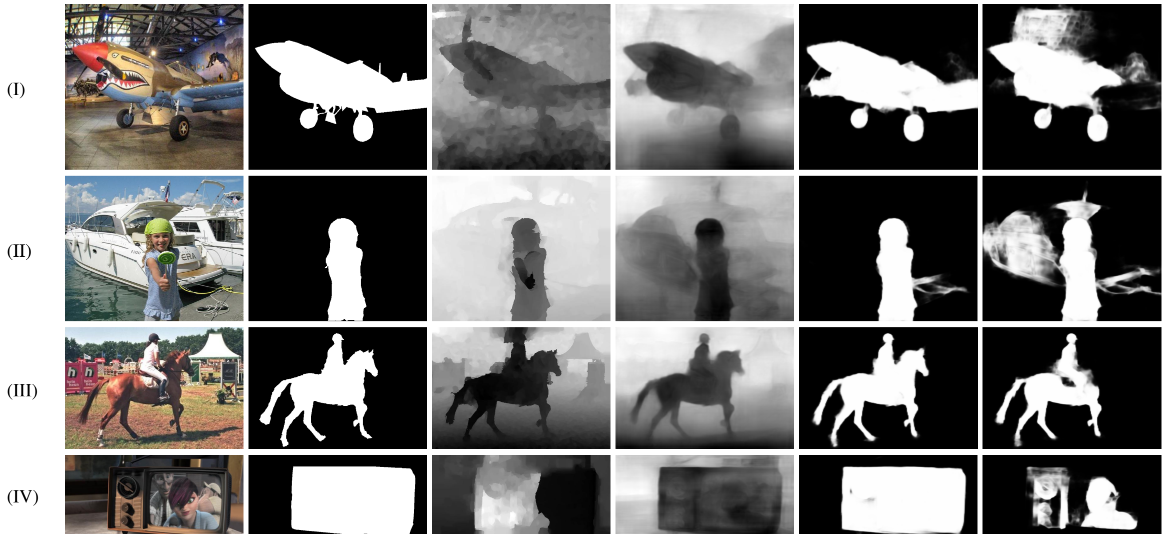
The quantitative comparison results are shown in Table I and visual comparisons are shown in Fig 5. *We can see that our model outperforms state-of-the-art RGB saliency models on five out of seven datasets. The comparison in terms of PR curves on four datasets are also given in Figure 6. These results demonstrate the effectiveness of the depth features we used and the importance of introducing depth information to deep saliency detection on these datasets. More surprisingly, our model also achieves comparable performance compared with contemporary RGB-D saliency detection methods (i.e., SSF, UCNet, and JLDCF) and outperforms previous ones on four datasets, i.e., NLPR, SSD, RGBD135, and STEREO. These results show the potential of the strategy that only involves depth data in training and without using it in testing for saliency detection. Aiming to achieve this same goal, our model outperforms the contemporary A2dele model on four datasets, i.e., NJUD, SSD, RGBD135, and STEREO, although A2dele also uses the training set of DUT-RGBD for training. Such a comparison clearly demonstrates the effectiveness of our mechanism to leverage the depth data.*

In Figure 5, we show some visual comparison between our model and state-of-the-art RGB and RGB-D saliency

TABLE II

Ablation study experimental results. *We conduct ablation study on the NJUD, NLPR, RGBD135, and SSD datasets.* Row (a) means the baseline U-Net trained by only RGB images. Row (b) means we add the depth branch and fuse the two kinds of decoding features. Row (c) means we further use DASPP and the proposed DMSF, but without using the NL model. Row (d) means we further use the NL model in DMSF. Red indicates the best performance.

Settings	NJUD [54]				NLPR [55]				SSD [56]				RGBD135 [44]			
	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
(a) RGB U-Net (baseline)	0.865	0.852	0.902	0.072	0.897	0.873	0.941	0.039	0.828	0.796	0.890	0.080	0.875	0.834	0.927	0.046
(b) +Depth	0.880	0.876	0.918	0.054	0.901	0.877	0.935	0.036	0.843	0.810	0.887	0.052	0.892	0.866	0.921	0.031
(c) +DMSF_w/o_NL	0.885	0.876	0.921	0.052	0.905	0.880	0.940	0.036	0.859	0.822	0.903	0.048	0.896	0.874	0.936	0.031
(d) +NL (Ours)	0.886	0.876	0.927	0.050	0.906	0.882	0.936	0.038	0.861	0.832	0.917	0.049	0.906	0.886	0.943	0.027



(a)Image (b)GT Saliency (c)GT Depth (d)Predicted Depth (e)+Depth (f)RGB U-Net

Fig. 7. **Visual comparison between ”RGB U-Net” and the ”+Depth” setting.** The GT depth maps and our predicted ones are also given.

models. We can see that our model not only outperforms RGB saliency models, but also can achieve comparable results *compared with* RGB-D saliency models. It can leverage depth cues to more accurately localize salient objects and ignore the disturbance from background objects by comparing the depths of the salient objects and other background ones. We can also see that our model can work well on various scenes, such as images with both simple and cluttered backgrounds, cartoon films, both indoor and outdoor scenes, which well demonstrates its robustness.

D. Ablation Study

To understand why our model performs well, we conduct ablation study experiments on *four datasets, i.e., NJUD [54], NLPR [55], RGBD135 [44] and SSD [56].* The qualitative results can be found in Table II. Row (a) means we train the baseline U-Net [34] architecture by only using RGB images of the two datasets. Row (b) means we add the depth branch and fuse its features with the RGB saliency branch using fusion decoding modules. Row (c) means we further use DASPP for the depth branch and also use the proposed DMSF module for the RGB saliency branch to fuse the depth DASPP features, but without using the NL model. The last row means we further use the NL model in DMSF, i.e., our whole network.

From the results, we can see that adding the depth branch and fusing its features for saliency detection can largely improve the model performance, especially for the MAE metric. Moreover, using the proposed DMSF model to fuse the multiscale DASPP features can bring further performance gains, especially on the *SSD* dataset. Finally, we can obtain the best performance *on three out of four datasets* by adding the NL module in DMSF to further incorporate global contexts. These results clearly demonstrate the effectiveness of our proposed ideas.

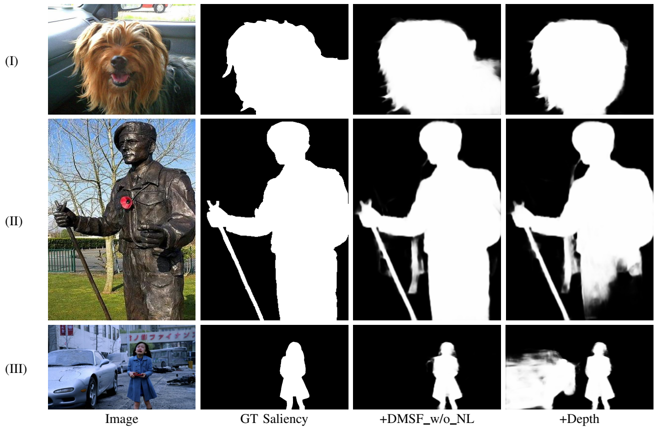


Image GT Saliency +DMSF_w/o_NL +Depth

Fig. 8. **Visual comparison between the “+DMSF_w/o_NL” and the “+Depth” setting.**

We also give qualitative results to show how our proposed

TABLE III

QUANTITATIVE COMPARISON AMONG OUR PROPOSED MODEL, BASELINE RGB U-NET, AND STATE-OF-THE-ART RGB SALIENT OBJECT DETECTION MODELS ON SIX RGB SALIENCY DATASETS. BLUE INDICATES THE BEST PERFORMANCE IN EACH GROUP.

	SOD [75]				DUT-O [20]				DUTS-TE [59]				ECSSD [76]				HKU-IS [25]				PASCAL-S [77]			
	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE	S_m	maxF	E_ξ	MAE
Amulet [32]	0.755	0.808	0.812	0.145	0.781	0.743	0.834	0.098	0.803	0.778	0.851	0.085	0.894	0.915	0.932	0.059	0.883	0.896	0.933	0.052	0.821	0.857	0.862	0.103
DSS [3]	0.741	0.847	0.813	0.128	0.788	0.771	0.845	0.066	0.822	0.825	0.884	0.057	0.882	0.921	0.931	0.052	0.880	0.913	0.938	0.040	0.774	0.849	0.860	0.113
BMP [62]	0.784	0.856	0.847	0.112	0.809	0.774	0.848	0.064	0.861	0.851	0.907	0.049	0.911	0.928	0.944	0.045	0.907	0.921	0.950	0.039	0.831	0.877	0.892	0.086
PiCANet [4]	0.787	0.855	0.846	0.108	0.826	0.794	0.865	0.068	0.861	0.851	0.915	0.054	0.914	0.931	0.953	0.047	0.906	0.921	0.951	0.042	0.837	0.880	0.900	0.088
R3Net [63]	0.761	0.816	0.835	0.124	0.817	0.760	0.857	0.063	0.835	0.801	0.881	0.057	0.910	0.926	0.949	0.040	0.895	0.904	0.944	0.036	0.807	0.800	0.853	0.092
CPD [64]	0.765	0.853	0.849	0.119	0.818	0.794	0.868	0.057	0.866	0.864	0.914	0.043	0.910	0.936	0.951	0.040	0.904	0.924	0.950	0.033	0.824	0.880	0.891	0.087
EGNet [6]	0.807	0.844	0.873	0.097	0.841	0.777	0.878	0.053	0.887	0.866	0.927	0.039	0.925	0.936	0.955	0.037	0.918	0.923	0.956	0.031	0.852	0.841	0.892	0.074
MINet [65]	0.805	0.836	0.870	0.092	0.833	0.769	0.869	0.056	0.884	0.865	0.927	0.037	0.925	0.938	0.957	0.033	0.919	0.926	0.960	0.029	0.856	0.846	0.903	0.064
ITSD [66]	0.809	0.844	0.873	0.093	0.840	0.792	0.880	0.061	0.885	0.868	0.929	0.041	0.925	0.939	0.959	0.034	0.917	0.926	0.960	0.031	0.859	0.855	0.908	0.066
RGB U-Net	0.786	0.811	0.857	0.099	0.821	0.753	0.856	0.065	0.862	0.831	0.906	0.050	0.911	0.920	0.946	0.044	0.901	0.907	0.946	0.039	0.849	0.839	0.897	0.073
Ours	0.795	0.814	0.867	0.094	0.820	0.747	0.848	0.069	0.861	0.824	0.899	0.052	0.914	0.921	0.946	0.044	0.904	0.906	0.944	0.039	0.847	0.833	0.893	0.076

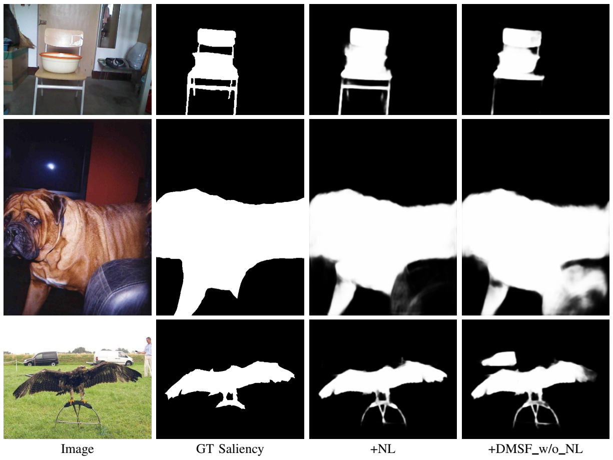


Fig. 9. **Visual comparison between the "+NL" and the "+DMSF_w/o_NL" setting.**

model improves performance.

In Figure 7 we show the comparison of "RGB U-Net" and the "+Depth" setting. We can see that adding the depth cues can help our saliency model removing the distraction from backgrounds (rows (I) and (II)) or recovering the missing parts of salient objects (rows (III) and (IV)). We also show the GT depth maps and our predicted depth maps in columns (c) and (d). We can see that the depth information supplies complementary cues with effective discrimination. We also find that sometimes the GT depth maps are noisy but our model can estimate more accurate depth (rows (I) and (IV)). This may be the reason why our model can sometimes outperform state-of-the-art RGB-D saliency models.

We also show the visual improvements of the "+DMSF_w/o_NL" and the "+NL" settings in Figure 8 and 9, respectively. The comparisons show that using the DMSF model and the NL branch can further help to discriminate and uniformly highlight the salient objects, thus demonstrating their effectiveness.

E. Discussion

In this section, we discuss whether our model can improve the RGB saliency detection performance and its model limitation.

1) Model performance on RGB saliency datasets: Since our model only requires RGB images as inputs during testing, it naturally raises a question that whether it can improve the RGB saliency detection performance. To answer this question, we compare our model with state-of-the-art RGB saliency methods, as well as the baseline U-Net model that does not involve depth data in training. The results are given in Table III. We can observe that our model can not outperform SOTA RGB saliency models on RGB saliency datasets. Compared with the baseline RGB U-Net model, our model shows better results on only two datasets, i.e., SOD and ECSSD. The probable reasons are two folds. First, RGB SOD has drawn extensive research interests for several years and many models resort to various elaborately designed methods to achieve precise saliency detection results, such as attention models, recurrent models, and complementary contour/edge features. In contrast, we only incorporate depth estimation into a U-Net model. Second, current RGB saliency datasets and RGB-D ones have different data distribution and properties. Depth cues may be more important for current RGB-D saliency datasets while it can not supply much informative cues for current RGB saliency datasets. Hence, the effectiveness of our proposed model depends on specific scenes. Not all visual scenes are suitable to use our proposed model. So do other SOTA saliency methods.

2) Relation between the depth estimation performance and saliency detection performance: Another two important questions our model raises are how is our depth estimation performance and what is the relation between the depth estimation performance and saliency detection performance. To answer the first question, we report the depth estimation performance in Table IV. Here we only conduct performance evaluation on the NLPR and RGBD135 datasets since they have the most accurate ground truth depth maps that are captured by Microsoft Kinect, while other datasets usually have very coarse depth maps. We adopt the widely used mean relative error (rel), root mean squared error (rms), mean log10 error (log10), and the accuracy under three thresholds, i.e., $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$. To answer the second question, we report the pearson correlation coefficient (PCC) between the saliency metric S_m and the depth accuracy

TABLE IV
DEPTH ESTIMATION PERFORMANCE AND THE PEARSON CORRELATION COEFFICIENT (PCC) BETWEEN THE DEPTH ESTIMATION PERFORMANCE AND SALIENCY DETECTION PERFORMANCE.

Dataset	Error (lower is better)			Accuracy (higher is better)			PCC
	rel	rms	log10	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	
NLPR [55]	1.811	0.272	0.294	0.333	0.518	0.639	0.051
RGBD135 [44]	1.012	0.267	0.266	0.244	0.463	0.631	-0.099

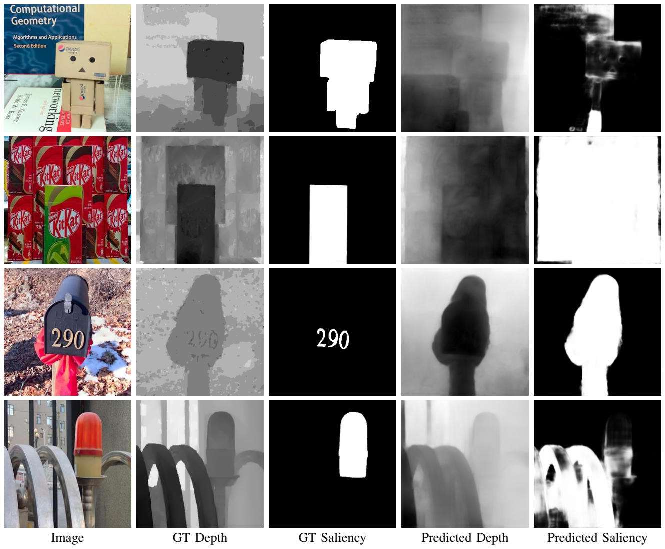


Fig. 10. **Failure case analysis.**

$\delta < 1.25^3$.

From Table IV, we find that our model does not achieve accurate depth estimation results and the correlation between the depth estimation performance and saliency detection performance is very weak. The probable reasons are three-folds. First, the ground truth depth maps in one of our training set, i.e., NJUD, are very coarse, hence leading to our coarse depth estimation results. Second, since saliency detection focuses more on relative depth relationships instead of accurate depth values, we adopt the depth ranking loss in [53] to train our depth branch. Hence, our model is not good at estimating accurate depth values. Third, our model performs multi-task learning but focuses more on saliency detection. Depth estimation can provide informative cues for saliency detection, while it has no proofs that saliency detection can benefit depth estimation. Hence, our proposed model naturally may downgrade the depth estimation performance. However, coarse depth information can also supply sufficient supplementary cues about the scene layout for saliency detection, thus it is not necessary to obtain very accurate depth estimation results for our model. This is also proved by the widely existing coarse depth maps in current RGB-D saliency detection benchmark datasets.

3) Model limitations: As aforementioned, our model fails to bring performance gains for current RGB saliency detection datasets. As for RGB-D saliency datasets, it also can not outperform state-of-the-art RGB-D saliency methods. However, it allows to infer RGB-D saliency without requiring depth input.

We also show some failure cases in Figure 10. We find that our model mainly fails because of two cases. The first is that the model fails to predict accurate depth maps hence resulting to incorrect saliency maps, as shown in the first two rows in Figure 10. The second is that the model predicts accurate depth maps but fails to intelligently combine both depth and appearance cues, hence also resulting to incorrect saliency detection results, as shown in the last two rows.

V. CONCLUSION

Depth information plays a very important role in the visual attention mechanism. However, directly collecting depth data for each image or video is expensive and impractical. In this paper, we have proposed to simultaneously estimate depth and detect saliency for RGB images in a unified deep CNN. Intermediate depth features can be fused with RGB saliency features to supply complementary information for improving the saliency detection performance. We further proposed to fuse multiscale depth and RGB features and also introduced global contexts. Experimental results clearly demonstrated the effectiveness of our proposed model, compared with both state-of-the-art RGB and RGB-D saliency models. We hope our work can inspire further research on leveraging depth cues for RGB saliency detection.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
[2] N. Liu and J. Han, “Dhsnet: Deep hierarchical saliency network for salient object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 678–686.
[3] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, “Deeply supervised salient object detection with short connections,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5300–5309.
[4] N. Liu, J. Han, and M.-H. Yang, “Picanet: Learning pixel-wise contextual attention for saliency detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3089–3098.
[5] M. Feng, H. Lu, and E. Ding, “Attentive feedback network for boundary-aware salient object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1623–1632.
[6] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, “Egnet: Edge guidance network for salient object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8779–8788.
[7] W. Z. J. Lin X and M. L, “Saliency detection via multi-scale global cues[j]. ieee transactions on multimedia,” *IEEE Transactions on Multimedia*, vol. 21(7), pp. 1646–1659, 2019.
[8] R. Y. Xu M and W. Z, “Saliency detection in face videos: A data-driven approach,” *IEEE Transactions on Multimedia*, vol. 20(6), pp. 1335–1349, 2017.
[9] K. A. Bak C and E. E, “Spatio-temporal saliency networks for dynamic saliency prediction,” *IEEE Transactions on Multimedia*, vol. 20(7), pp. 1688–1698, 2017.

[10] L. W. Imamoglu N and F. Y, "A saliency detection model using low-level features based on wavelet transform," *IEEE Transactions on Multimedia*, vol. 15(1), pp. 96–105, 2013.

[11] 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 *European Conference on Computer Vision (ECCV)*. Springer, 2018.

[12] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion," *IEEE Transactions on Cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2017.

[13] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

[14] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *IEEE International Conference on Computer Vision*, 2019, pp. 7254–7263.

[15] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936.

[16] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3684–3692.

[17] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[18] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2010.

[19] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[20] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[21] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on markov absorption probabilities," *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1639–1649, 2015.

[22] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 12, pp. 2009–2021, 2013.

[23] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *IEEE International Conference on Computer Vision*, 2011, pp. 914–921.

[24] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[25] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.

[26] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[28] X. Fan, C. Xiang, C. Chen, P. Yang, L. Gong, X. Song, P. Nanda, and X. He, "Buildsensys: Reusing building sensing data for traffic prediction with cross-domain learning," *IEEE Transactions on Mobile Computing*, 2020.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[31] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6593–6601.

[32] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 202–211.

[33] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *European Conference on Computer Vision*, 2018, pp. 234–250.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[35] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[36] S. Xie and Z. Tu, "Holistically-nested edge detection," in *IEEE International Conference on Computer Vision*, 2015, pp. 1395–1403.

[37] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.

[38] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1711–1720.

[39] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "Capsal: Leveraging captioning to boost semantics for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6024–6033.

[40] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *European Conference on Computer Vision*, 2018, pp. 355–370.

[41] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1448–1457.

[42] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3917–3926.

[43] X. Xiao, Y. Zhou, and Y.-J. Gong, "Rgb-'d'saliency detection with pseudo depth," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2126–2139, 2018.

[44] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *International Conference on Internet Multimedia Computing and Service*. ACM, 2014, p. 23.

[45] A. Ciptadi, T. Hermans, and J. Rehg, "An in depth view of saliency," in *British Machine Vision Conference*, 2013.

[46] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "Rgbd salient object detection via deep fusion," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2274–2285, 2017.

[47] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient object detection for rgb-d image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, 2019.

[48] N. Wang and X. Gong, "Adaptive fusion for rgb-d salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.

[49] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765.

[50] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9060–9069.

[51] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[53] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 311–320.

[54] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *International Conference on Image Processing*. IEEE, 2014, pp. 1115–1119.

[55] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *European Conference on Computer Vision*. Springer, 2014, pp. 92–109.

[56] C. Zhu and G. Li, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *International Conference on Computer Vision Workshops*, 2017, pp. 3008–3014.

[57] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *IEEE International Conference on Computer Vision*, 2017, pp. 4558–4567.

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

[59] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, “Learning to detect salient objects with image-level supervision,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 136–145.

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[61] 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.

[62] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, “A bi-directional message passing model for salient object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1741–1750.

[63] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, “R3net: Recurrent residual refinement network for saliency detection,” in *International Joint Conference on Artificial Intelligence*, 2018, pp. 684–690.

[64] Z. Wu, L. Su, and Q. Huang, “Cascaded partial decoder for fast and accurate salient object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[65] Y. Pang, X. Zhao, L. Zhang, and H. Lu, “Multi-scale interactive network for salient object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9413–9422.

[66] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, “Interactive two-stream decoder for accurate and fast saliency detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9141–9150.

[67] 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.

[68] H. Chen and Y. Li, “Three-stream attention-aware network for rgb-d salient object detection,” *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[69] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, “Select, supplement and focus for rgb-d saliency detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3472–3481.

[70] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, “Uc-net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8582–8591.

[71] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, “Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3052–3062.

[72] Y. Niu, Y. Geng, X. Li, and F. Liu, “Leveraging stereopsis for saliency analysis,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 454–461.

[73] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, “Saliency detection on light field,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813.

[74] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, “Basnet: Boundary-aware salient object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[75] V. Movahedi and J. H. Elder, “Design and perceptual validation of performance measures for salient object segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 49–56.

[76] Q. Yan, L. Xu, J. Shi, and J. Jia, “Hierarchical saliency detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.

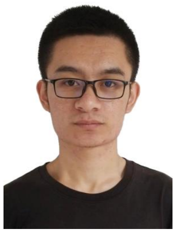
[77] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, “The secrets of salient object segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

Yuan-fang Zhang is currently pursuing his dual Ph.D. degrees with the School of Computer Science at Northwestern Polytechnical University, Shaanxi, China and Faculty of Engineering and IT, University of Technology Sydney, Australia. His current research focuses on image processing and computer vision, especially in the domains of the image enhancement, object detection and saliency detection.

Jiangbin Zheng received the Ph.D. degree from Northwestern Polytecnical University in 2002, where he is a Full Professor and Dean with School of Software. His research interests include computer graphics, computer vision and multimedia. He has published over 100 papers in the above related research area.

Wenjing Jia received her Ph.D. degree in Computing Science from the University of Technology Sydney (UTS) in 2007. She is currently a Senior Lecturer at the Faculty of Engineering and IT and a Core Research Member at the Global Big Data Technologies Centre, UTS. She has authored over 100 quality journal articles and conference papers. Her research interests include image/video analysis, computer vision, and pattern recognition.

Wenfeng Huang is now doing internship at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He has won the ACM-ICPC Asia Regional Contest Silver Medal and the ACM-ICPC Asia-East Continent Final Bronze Medal. His research interests include image enhancement, saliency detection, object detection and medical image processing.



Long Li is a M.E. student of School of Automation at Northwestern Polytechnical University, Xi’an, China. He received the B.E. degree from Northwestern Polytechnical University in 2018. His research interests include computer vision and machine learning.

Nian Liu is a postdoctoral researcher with Mohamed bin Zayed University of Artificial Intelligence. He received the Ph.D. degree and the B.S. degree from School of Automation at Northwestern Polytechnical University, in 2020 and 2012, respectively. His research interests include computer vision and machine learning, especially on saliency detection and deep learning.

Fei Li received his B.S. degree from the School of Software at Northwestern Polytechnical University in 2016.06. He is currently pursuing his Ph.D. degree with the School of Software at Northwestern Polytechnical University, Shaanxi, China. His current research focuses on image processing and computer vision, particularly in the domains of deep learning-based underwater image restoration and enhancement.

Xiangjian He received the Ph.D. degree in Computer Science from the University of Technology Sydney (UTS), Australia in 1999. He is currently a Full Professor and the Director of the Computer Vision and Pattern Recognition Laboratory, Global Big Data Technologies Centre, UTS.