Boosting RGB-D Saliency Detection by Leveraging Unlabeled RGB Images

Xiaqiang Wang, Lei Zhu, Siliang Tang, Huazhu Fu, Senior Member, IEEE, Ping Li, Fei Wu, Senior Member, IEEE, Yi Yang, Senior Member, IEEE, Yueting Zhuang, Senior Member, IEEE

Abstract—Training deep models for RGB-D salient object detection (SOD) often requires a large number of labeled RGB-D images. However, RGB-D data is not easily acquired, which limits the development of RGB-D SOD techniques. To alleviate this issue, we present a Dual-Semi RGB-D Salient Object Detection Network (DS-Net) to leverage unlabeled RGB images for boosting RGB-D saliency detection. We first devise a depth decoupling convolutional neural network (DDCNN), which contains a depth estimation branch and a saliency detection branch. The depth estimation branch is trained with RGB-D images and then used to estimate the pseudo depth maps for all unlabeled RGB images to form the paired data. The saliency detection branch is used to fuse the RGB feature and depth feature to predict the RGB-D saliency. Then, the whole DDCNN is assigned as the backbone in a teacher-student framework for semi-supervised learning. Moreover, we also introduce a consistency loss on the intermediate attention and saliency maps for the unlabeled data, as well as a supervised depth and saliency loss for labeled data. Experimental results on seven widely-used benchmark datasets demonstrate that our DDCNN outperforms state-of-the-art methods both quantitatively and qualitatively. We also demonstrate that our DDCNN can further improve the performance, even when using an RGB image with the pseudo depth map.

Index Terms—RGB-D salient object detection, semi-supervised learning, depth estimation and attention consistency.

I. INTRODUCTION

RGB-D salient object detection (SOD) has attracted a surge in interest recently [1], [2], [3]. Early RGB-D detectors [4], [5], [6], [7] mainly examined the handcrafted priors, which degrades the detection performance since the assumptions of these heuristic priors are not always correct. More recently, RGB-D SOD detectors [8], [9], [10], [11], [12], [13], [14], [15] based on convolutional neural networks (CNNs) have been developed by learning the features from RGB images and depth maps and exploring the complementary information between them. Although these models have achieved impressive performances on the benchmark datasets, there are still several issues limiting the development of RGB-D SOD techniques: 1) pixel-level annotation for supervised learning is expensive and time-consuming, and 2) compared to RGB images, paired RGB-D images are more difficult to collect. Fortunately, it is easy to collect a large number of unlabeled RGB images. Thus, how to leverage the unlabeled RGB images to assist the RGB-D SOD methods is a desirable direction to explore.

In this paper, we present a novel dual-semi RGB-D salient object detection network (DS-Net) for boosting the RGB-D saliency detection by leveraging unlabeled RGB data, as shown in Fig. 1. Our DS-Net has a two-fold motivation: semi-supervised learning with unlabeled data and semi-paired data including RGB images without depth maps. Specifically, we devise a depth decoupling convolutional neural network (DDCNN) to estimate depth maps of RGB images and detect RGB-D saliency maps jointly. Our DDCNN disentangles two types of features from the RGB image, i.e., depth-aware features and depth-dispelled features. The depth-aware features are used to estimate the pseudo depth maps for RGB images, while the depth-dispelled features are extracted from the input RGB image and then fused with the depth map features to predict the RGB-D saliency. Finally, we embed DDCNN as the backbone into a teacher-student framework to provide semi-supervision for training the whole DS-Net based on labeled RGB-D data and unlabeled RGB images with their pseudo depth maps. Moreover, we also introduce a consistency loss to constrain the attention and saliency maps on unlabeled data, assisting with the supervised loss for saliency and depth predictions on labeled data. In summary, the main contributions are:

- A dual-semi RGB-D salient object detection network (DS-Net) is proposed for leveraging RGB images as unlabeled data to assist the RGB-D SOD task in a semi-supervised manner. We show an effective solution to improve the RGB-D task performance by incorporating RGB images with the pseudo depth maps.
- A depth decoupling convolutional neural network (DDCNN) is designed to jointly estimate depth maps for RGB images and predict saliency maps for RGB-D images. Two types of features are disentangled from RGB image, i.e., depth-aware features and depth-dispelled features, enabling the network to identify the latent features specific to each modality and task.
- For enhancing the semi-supervised consistency, a consistency loss is introduced in the teacher-student network to constrain the intermediate attention and saliency maps on the unlabeled data, assisting with the supervised depth and saliency loss on labeled data.
- Last but not least, experimental results on seven widely-
used RGB-D SOD datasets show that our DDPCNN outperforms state-of-the-art methods. We also demonstrate that the semi-supervised DS-Net can further improve the performance, even when using an RGB image with the pseudo depth map.

Our code, the trained models, and the predicted saliency maps on all seven benchmark datasets are released at: https://github.com/Robert-xiaoqiang/DS-Net.

II. RELATED WORK

A. RGB-D Salient Object Detection

RGB-D SOD methods based on deep learning can be roughly grouped into three categories: early fusion, middle fusion, and later fusion. Early fusion concatenates the RGB and depth images as a four-channel input and then passes this into CNNs for saliency detection. Liu et al. [16], and Huang et al. [17] developed a single-stream convolutional neural network and a fully convolutional network (FCN) with short connections to detect salient regions from the concatenated four-channel input, respectively. Late fusion employs two separate backbone networks for RGB and depth to generate individual features which are fused together for final prediction. Han et al. [18] transferred the structure of the RGB-based CNN to be applicable for the depth view and fused the deep representations of both views automatically to obtain the final saliency map. Wang et al. [19] designed a two-streamed CNN to predict a saliency map from each modality separately and fuse the predicted saliency maps adaptively by learning a switch map.

As the most popular CNN framework for RGB-D saliency detection, middle fusion typically integrates multi-scale intermediate features from input RGB and depth modalities in different manners. Chen et al. [8] developed complementarity-aware fusion (CA-Fuse) modules to progressively integrate RGB features and depth features. Piao et al. [10] devised depth refinement blocks to extract and fuse multi-level paired complementary RGB and depth cues. Fan et al. [3] built a depth-depurator network to filter out noise in the depth map for better fusing cross-modal features. Fu et al. [11] utilized a shared backbone to extract hierarchical features from RGB and depth inputs simultaneously for a multi-scale cross-module fusion. Zhang et al. [12] presented a probabilistic RGB-D saliency detection network via conditional variational autoencoders to approximate human annotation uncertainty and produce multiple saliency maps for each input image. Zhang et al. [13] included complementary interaction models, consisting of a cross-modal attention unit and a boundary supplement unit, to select useful RGB and depth features for salient object location and boundary detail refinement. Li et al. [20] adopted a three-level Siamese encoder-decoder structure to develop three modules to fuse low-level, middle-level, and high-level RGB and depth features, respectively, for cross-modal and cross-scale RGB-depth interactions.

Although existing CNN-based methods have achieved more accurate results than traditional RGB-D saliency detectors, their network training requires a large amount of data with pixel-level saliency annotations. Moreover, annotated training data are collected from limited scenarios, causing the networks to suffer from degraded performance on unseen photos. As such, this work presents a semi-supervised network to fuse unlabeled data with the labeled data for boosting RGB-D saliency detection. More importantly, rather than relying on unlabeled RGB-D paired images, our unlabeled data consists of only RGB images, which are much easier to collect in our daily life.

B. Semi-Supervised Learning

By integrating labeled and unlabeled data for network training, semi-supervised learning (SSL) has achieved remarkable results in many computer vision tasks [21], [22]. As a typical kind of SSL technique, self-ensembling usually devises a consistency loss on the unlabeled data to guarantee invariant predictions for perturbations of unlabeled data. For example, the Π-model [23] devised consistency constraints between the current network prediction and the temporal average of network predictions for unlabeled data. The mean teacher (MT) framework [24] proposed to ensemble the network parameters to replace the network predictions of the Π-model [23], achieving improved performance in the semi-supervised learning. Developing semi-supervised CNNs for RGB-D saliency detection usually requires numerous paired RGB-D images, which are not easy to collect. RGB images, however, are much easier to collect. Hence, our model is
devised to leverage unpaired RGB images to formulate semi-supervised RGB-D saliency detection.

III. METHODOLOGY

Fig. 1 shows an illustration of our dual-semi RGB-D salient object detection network (DS-Net), which integrates the labeled RGB-D data and unlabeled RGB images. A depth decoupling convolutional neural network (DDCNN) is utilized as the backbone, which contains two components: a depth estimation branch and a saliency detection branch for estimating the depth maps of RGB images and predicting the saliency maps of RGB-D images, respectively. To train the whole DS-Net using labeled RGB-D data and unlabeled RGB images, we first train the depth estimation branch of DDCNN using labeled RGB-D data to learn the mapping from an RGB image to its depth map. Then, we estimate the pseudo depth maps for all unlabeled RGB images to form the paired data. Finally, we utilize DDCNN as the backbone in a teacher-student framework. To train the whole DS-Net in a semi-supervised manner, we utilize a supervised loss on depth and a consistency loss on intermediate attention maps and saliency predictions for all unlabeled RGB images to form the paired data.

In the depth estimation branch of DDCNN, each RGB feature $R_i$ is disentangled into a depth-aware feature $R_i^d$ for estimating the depth map and a depth-dispersed feature $R_i^s$ for predicting saliency, using a convolutional block with “Conv(3×3) → BN → RelU → Conv(3×3)” The depth-aware features $R_i^d$ are then up-sampled to the same spatial resolution of $R_i^d$ and concatenated together to predict a depth map by applying a “Conv(3×3) → Conv(1×1)” convolutional block. Moreover, we fuse $R_i^d$ and $R_i^s$ to reconstruct RGB features $\tilde{R}_i$ and compute a reconstruction loss $L_i^r$ to regularize the decoupling process:

$$L_i^r = L_{MSE}(\tilde{R}_i, R_i),$$

where $CB(\cdot)$ is a convolutional block with “Conv(3×3) → BN → RelU → Conv(3×3)”. $Cat(\cdot)$ denotes a feature concatenation operation, and $L_{MSE}$ represents the mean square error (MSE) loss.

In the saliency detection branch of DDCNN, we devise a depth-induced fusion module (DIM) to fuse depth-dispersed features $R_i^s$ with two depth features ($R_i^d$ and $D_i$) at each CNN layer. The fused features from DIMs at the four levels of feature extraction layers are denoted as $\{F_i\}_{i=1}^4$, as shown in the yellow block of Fig. 2. Finally, we iteratively merge the four obtained features $\{F_i\}_{i=1}^4$ to predict a saliency map with a 3×3 convolution, a 1×1 convolution, and a sigmoid function.

A. DDCNN

Fig. 2 illustrates the architecture of our DDCNN, which is a two-branch structure including a depth estimation branch (green block) and saliency detection branch (yellow block).
When merging features \((F_i, F_{i+1})\) from two adjacent CNN layers, we up-sample the low-resolution features \(F_{i+1}\) to the same resolution as the high-resolution features \(F_i\), and then refine the feature map \(F_i\) from the shallow layer by passing it to an atrous spatial pyramid pooling (ASPP) module [25] with four dilated convolutional layers (dilation rates: 1, 6, 12, 18), followed by a \(1 \times 1\) convolutional layer on the four dilated features concatenated together. We then fuse the refined features of ASPP with up-sampled \(F_{i+1}\) to produce the merged features of \(F_i\) and \(F_{i+1}\).

**Depth-Induced Fusion Module.** As shown in Fig. 2, a DIM at each CNN layer merges three feature maps (i.e., \(R_i^d\), \(R_i^s\), and \(D_i\)) to leverage the complementary information of RGB and depth modalities. Fig. 3 gives an illustration of the DIM at the \(i\)-th CNN layer, which takes depth-aware feature \(R_i^d\), depth-dispersed feature \(R_i^s\), and depth feature \(D_i\) as inputs and produces a fused feature \(F_i\). To be specific, we first devise a depth-gated module (DGM) to fuse depth features \(D_i\) from the input depth map and \(R_i^d\) from the depth estimation branch. The fused features of DGM are denoted as \(F_{i}^{dgm}\). Then, we adopt a depth-awareness module (DAM) to fuse \(R_i^s\) and depth features \(D_i\) from the input depth map to obtain new features \(F_i^{dam}\). Moreover, we concatenate \(F_i^{dgm}\) and \(F_i^{dam}\), and apply a \(3 \times 3\) convolutional layer and a sigmoid function on the concatenated features to learn an attention map \(A_i^d\) for weighting depth features \(D_i\), thereby generating a new feature map \(F_i^d\). Finally, the output of DIM (denoted as \(F_i\)) is computed by adding \(F_i^{dam}\), \(F_i^{dgm}\), and \(F_i^d\). 

\[
F_i = F_i^{dam} + F_i^{dgm} + F_i^d ,
\]

**Depth-Awareness Module.** Considering the redundancy and noise in low-quality depth maps and the intrinsic difference between RGB and depth features, we design an effective fusion method to suppress noise and utilize the complementary information from the features \(R_i^s\) and \(D_i\) of the two modalities. Inspired by CBAM [26], we design a DAM equipped with a channel attention and a spatial attention operation to merge RGB features \(R_i^s\), and depth features \(D_i\). Fig. 3 shows the workflow of DAM, which outputs a new feature map \(F_i^{dam}\). This is achieved by applying a channel attention operation on \(D_i\) to weight different channels of \(R_i^s\), and then a spatial attention is computed on \(D_i\) to recalibrate pixel-wise saliency cues of \(R_i^s\) to obtain \(F_i^{dam}\). 

\[
F_i^{dam} = S_{att}(D_i) \odot (C_{att}(D_i) \odot R_i^s) ,
\]

where the channel attention \(C_{att}()\) includes a \(3 \times 3\) convolution and a global average pooling. The spatial attention \(S_{att}()\) consists of a \(3 \times 3\) convolution with a softmax function. \(\odot\) denotes a channel-wise multiplication while \(\odot\) represents an element-wise multiplication.

**Depth-Gated Module.** \(R_i^d\) are the depth signals from RGB image, hence combining them and depth features \(D_i\) from the input depth image enriches the depth representation from different depth modalities. As such, we devise a DGM to merge \(R_i^d\) (size: \(H \times W \times C\)) and \(D_i\) (size: \(H \times W \times C\)) by considering long-range pixel dependencies for learning saliency cues. First, we first apply a \(3 \times 3\) convolutional layer on \(R_i^d\), reshape the resultant features as an intermediate \(C \times HW\) feature map, apply another \(3 \times 3\) convolutional layer on \(R_i^d\), reshape the resultant features as another intermediate \(HW \times C\) feature map, and then multiply two intermediate features to generate a non-local similarity matrix with size \(HW \times HW\). Furthermore, we apply a \(3 \times 3\) convolutional layer on \(D_i\) and reshape the resultant features as an intermediate \(HW \times C\) feature map, which is then multiplied with the non-local similarity matrix. After that, we reshape the features \(HW \times C\) resulting from the multiplication to a feature map with size \(H \times W \times C\), which undergoes a \(3 \times 3\) convolution to obtain the features \(F_i^{dgm}\) of DGM.

**B. Training Strategies**

**Supervised loss for labeled data.** For labeled data, we have a pair of input RGB and depth images with the corresponding annotated saliency mask. It is natural to take the annotated saliency mask as the ground truth \((G_s)\) for RGB-D saliency detection. On the other hand, we also have the ground truth \((G_d)\) for the depth estimation task, which is the input depth map. With the two ground truths \((G_s, G_d)\), the supervised...
TABLE I: Comparisons our network against state-of-the-art detectors on seven benchmark datasets. The best results are highlighted in bold. DS-Net (HRNet as backbone) is marked in bold when it outperforms all other baselines.

| Dataset | Metric | FL | YOLOv3 | SSDP | ATSA | LCF | DSA | DPP | DSC | HRNet | DMRA | ATSA | SSDP | DSA | DS-Net |
|---------|--------|----|-------|------|------|-----|-----|-----|-----|-------|------|------|------|-----|--------|
| NUI2K   | 0.881  | 0.837 | 0.837 | 0.888 | 0.888 | 0.837 | 0.883 | 0.837 | 0.888 | 0.888 | 0.888 | 0.888 | 0.888 | 0.888 | 0.888 | **bold** |
| NLPR    | 0.938  | 0.942 | 0.942 | 0.939 | 0.939 | 0.942 | 0.942 | 0.942 | 0.942 | 0.942 | 0.942 | 0.942 | 0.942 | 0.942 | 0.942 | **bold** |
| STERE   | 0.919  | 0.934 | 0.934 | 0.927 | 0.927 | 0.934 | 0.934 | 0.934 | 0.934 | 0.934 | 0.934 | 0.934 | 0.934 | 0.934 | 0.934 | **bold** |
| RGBD135 | 0.856  | 0.855 | 0.855 | 0.857 | 0.857 | 0.855 | 0.855 | 0.855 | 0.855 | 0.855 | 0.855 | 0.855 | 0.855 | 0.855 | 0.855 | **bold** |
| LFSF    | 0.698  | 0.695 | 0.695 | 0.702 | 0.702 | 0.695 | 0.695 | 0.695 | 0.695 | 0.695 | 0.695 | 0.695 | 0.695 | 0.695 | 0.695 | **bold** |
| SIP     | 0.944  | 0.947 | 0.947 | 0.951 | 0.951 | 0.947 | 0.947 | 0.947 | 0.947 | 0.947 | 0.947 | 0.947 | 0.947 | 0.947 | 0.947 | **bold** |

TABLE II: Quantitative results of our method and baseline networks for the ablation experiments on NJU2K [40], NLPR [41], STERE [42], RGBD135 [43], LFSF [44], and SIP [3]. “SL” denotes supervised learning, while “DS” represents semi-supervised learning.

| Name Networks | SL | SSL | NJU2K [40] | NLPR [41] | STERE [42] | RGBD135 [43] | LFSF [44] | SIP [3] |
|---------------|----|-----|------------|-----------|------------|-------------|----------|--------|
| M1            | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |
| M2            | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |
| M3            | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |
| M4            | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |
| M5            | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |
| M6            | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |
| M7            | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |
| M8            | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |
| M9            | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |
| M10           | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |

Our DS-Net | 0.981 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 | 0.983 |

loss (denoted as $L^c$) for a labeled image ($x$) is computed as the summation of the saliency detection loss and depth estimation loss:

$$L^c(x) = L_{BCE}(P_x, G_x) + \alpha L_{MSE}(P_d, G_d),$$

(4)

where $P_x$ and $P_d$ denote the predicted saliency map and depth map, respectively. $L_{BCE}$ and $L_{MSE}$ are the binary cross-entropy (BCE) loss and MSE loss functions, respectively. We empirically set the weight $\alpha = 1.0$ during the network training.

**Consistency loss for unlabeled data.** As a well-known semi-supervised manner, the teacher-student approach ensembles parameters of the network at different training stages as teacher network’s parameters and devises a consistency loss to make the student network learn from the teacher network for improving the quality of the network predictions. In detail, we feed the unlabeled RGB data into the depth estimation branch of DCCNN to predict a depth map and use this prediction as a pseudo label of the depth map for the unlabeled RGB data. After that, we pass the depth map and corresponding pseudo depth image into the student and teacher networks to obtain two groups of depth map and saliency map prediction results. We then fuse the predicted depth map and the four learned attention maps of DCCNN to be consistent for the student and teacher networks. A consistency loss $L^c$ on unlabeled data $y$ is defined as:

$$L^c(y) = L_{MSE}(S_y, T_y) + \sum_{l=1}^{N} L_{MSE}(S^l_y, T^l_y),$$

(5)

where $S_y$ and $T_y$ denote two detected saliency maps of the student and the teacher networks. $S^l_y$ and $T^l_y$ represent the learned attention map of DCCNN at the $l$-th CNN layer from the student and the teacher networks. We empirically set $\gamma = 0.1$ in this work.

**Overall loss for the student network.** The total loss of our network is computed as:

$$L_{total} = \frac{1}{N_1} \sum_{i=1}^{N_1} L^c(x_i) + \beta 1 \sum_{l=1}^{N} L_{MSE}(S^l_y, T^l_y) + \lambda \sum_{j=1}^{N} L_{MSE}(y_j, T^l_j)(y_j),$$

(6)

where $N_1$ and $N_2$ are the number of labeled RGB-D image pairs and unlabeled RGB images in our training set. $L^c(x_i)$ denotes the supervised loss (Eq. (4)) for the $i$-th labeled image $x_i$ while $L^c(y_i)$ is the consistency loss (Eq. (5)) for the $j$-th unlabeled image $y_j$. $L^c(S^l_i, T^l_i)$ and $L^c(y_j, T^l_j)$ represent...
the reconstruction loss (Eq. (1)) for labeled $x_i$ and unlabeled $y_j$, respectively. We empirically set the weights $\beta_1 = 0.01$ and $\beta_2 = 1.0$. The weight $\lambda$ balances the loss between the labeled and unlabeled data. Following [21], we use a time dependent Gaussian warm-up function to update $\lambda$: $\lambda(t) = \lambda_{\text{max}} e^{-(5(1-t/t_{\text{max}})^2)}$, where $t$ denotes the current training iteration and $t_{\text{max}}$ is the maximum training iteration. In our experiments, we empirically set $\lambda_{\text{max}} = 1.0$.

Following existing self-ensembling frameworks [24], we minimize the total loss $L_{\text{total}}$ of Eq. (6) to train the student network. The parameters of the teacher network are computed as the exponential moving average (EMA) of the parameters of the student network to ensemble the information in different training steps. Following [21], during the inference stage, given an input RGB image and an input depth image, we pass them into the student network to predict a saliency detection map, which is then taken as the final result of our semi-supervised RGB-D saliency detection network, and the teacher network is not involved in the inference stage.

Training configurations. We use the same HRNet architecture as the backbone for both the RGB image and the depth map. Given a 256x256 input image, the HRNet outputs multi-scale features at four convolutional blocks (layer 8, 27, 63, and 104) with a resolution of 64x64, 32x32, 16x16, and 8x8, respectively. These features are fed straight into the following network pipeline. As for the VGG-16, VGG-19 and ResNet-50 backbones, we employ the outputs from the layer (4, 7, 10, 13), (4, 8, 12, 16), and (11, 23, 41, 50) with the same resolution as the HRNet, respectively.

IV. EXPERIMENTAL RESULTS

Benchmark Datasets. We employ seven widely-used benchmark datasets to evaluate our network and state-of-the-art RGB-D saliency detectors. They are (i) NJU2K [40] (2,000 images), (ii) NLPR [41] (1,000 images), (iii) STERE [42] (1,000 images), (iv) RGBD135 [43] (135 images), (v) LFSR [44] (100 images), (vi) SIP [3] (929 images), (vii) DUTD [10] (1,200 images). On the DUTD, we follow the setting of [10] and use 800 images for training and 400 images for testing. For the other datasets, we follow recent works [9], [11], [3] to utilize the same 1,500 images from NJU2K and the same 700 images from NLPR as labeled RGB-D images to train our network for fair comparisons. For unlabeled RGB images, we utilize 10,553 RGB images from the training set of DUTS [45]. Although each training image of DUTS has the annotations of the saliency map, we do not use any saliency information of these images when training our dual-semi RGB-D saliency detector.

Evaluation metrics. We adopt four widely-used metrics to quantitatively compare RGB-D saliency detection...
performance, including S-measure ($S_m$) [46], F-measure ($F_\beta^{\max}$) [47], E-measure ($F_\phi^{\max}$) [48], and Mean Absolute Error (MAE) [49]. In general, a more accurate RGB-D saliency detector shall have a larger $S_m$, a larger $F_\beta^{\max}$, a larger $F_\phi^{\max}$, and a smaller MAE.

**Implementation Details.** During training, the backbone is initialize by popular backbones, such as VGG-16 [50], VGG-19 [50], ResNet-50 [51] and HRNet [52], which has been well-trained for the image classification task on the ImageNet [53] and other layers are randomly initialized. Training data is resized to 256 × 256 and augmented by random rotation and horizontal flipping. In addition, color jittering is used for the perturbation of unlabeled data. Stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0005 is used to optimize the whole network. The learning rate is adjusted by a poly strategy [54] with an initial learning rate of 0.001 and the power of 0.9. The whole training takes 25 hours with training batch size 8 (i.e., 4 labeled pairs and 4 unlabeled RGB images) and a maximum iteration of 20,000 on a single NVIDIA GTX 2080Ti GPU.

### A. Comparison with Baselines

We evaluate the effectiveness of our network by comparing it against 23 state-of-the-art RGB-D salient object detectors. They are LBE [4], DF [27], CTMF [18], PCF [8], TANet [28], CPFP [9], DMRA [10], D^2Net [3], SSF [13], UCNet [12], JLDCF [11], JLDCF[J] [29], HDF-Net [30], ATSA [35], SSDP [36], DSA^2F [37], PGA-Net [31], DANet [32], cmMS [33], Cas-Gnn [34], CMWNet [20], CoNet [39], and BBS-Net [38]. Note that LBE [4] utilizes handcrafted features to infer salient objects, while other 22 methods employ different deep networks to learn discriminative features for RGB-D saliency detection. To make the comparisons fair, we obtained the saliency maps of all baselines either from the authors or by using their released training models and parameters.

Table I reports the $S_m$, $F_\beta^{\max}$, $F_\phi^{\max}$, and MAE values of our method and baselines on all seven benchmark datasets. From the results, we find that DDCNN, *i.e.*, our network with only labeled data, also outperforms other compared RGB-D saliency detectors in terms of four evaluation metrics. By considering unlabeled RGB images, our DS-Net further improves the performance over other saliency detectors on almost all of the seven datasets. This indicates that our network
can more accurately identify salient objects from the input pair of RGB-D data than the compared detectors. Besides, we also explore the efficiency of semi-supervised learning without any extra RGB images. Specifically, we construct an additional experiment by splitting 3/10 labeled RGB-D data as the unlabeled RGB images (denoted as “DDCNN-semi-ourSplit”). As shown in Table I, “DDCNN-semi-ourSplit” still outperforms DMRA [10], but fails to suppress the most recent state-of-the-art supervised methods (i.e., BBS-Net [38] and CoNet [39]) due to the use of only 7/10 labels of RGB-D datasets. Compared to the best-performing results, our network achieves a $S_m$ improvement of 1.59\%, a $F_g^{\text{max}}$ improvement of 1.63\%, a $F_m^{\text{max}}$ improvement of 0.66\%, and a $\text{MAE}$ decrease of 9.89\% on the average of seven benchmarks.

Fig. 4 visually compares the saliency maps produced by our network and the state-of-the-art RGB-D saliency methods. By observing the different saliency maps, we can conclude that other compared methods in Fig. 4 (e)-(m) tend to include non-salient backgrounds or lose salient details in their predicted saliency maps, whereas our DS-Net better detects salient objects from input RGB-D image pairs and our results are more consistent with the ground truths (see Fig. 4 (c)). This indicates that, by leveraging only unlabeled RGB images, our network can suppress non-salient objects and detect more salient pixels than the state-of-the-art RGB-D saliency detectors, which are mainly trained in a supervised manner.

B. Ablation Analysis

Baseline Network. We perform ablation experiments to evaluate the effectiveness of the major components in our DS-Net. We consider ten baselines. The first seven employ supervised learning for RGB-D saliency detection, using DDCNN as the backbone, without the teacher network. Specifically, we firstly verify the contributions of fusion between RGB and depth features. “DDCNN-w/o-DAM”; “$M_1$”; “DDCNN-w/o-DGM”; “$M_2$” and “DDCNN-w/o-DIM”; “$M_3$” replace DAM, DGM and DIM with a simple feature concatenation operation, respectively. The fourth baseline (denoted as “DDCNN-w/o-depth”; “$M_4$”) removes the depth estimation branch from our DDCNN and does not separate RGB features but directly fuses them with depth features via DAM modules for saliency detection. The fifth baseline (denoted as “DDCNN-w/o-reconstruction-loss”; “$M_5$”) removes the reconstruction loss from our DDCNN. For the other two baselines with supervised learning, one (denoted as “DDCNN-w/o-pretrain”; “$M_6$”) directly trains DDCNN on labeled RGB-D data in an end-to-end manner, while the other (“$M_7$”) is our full DCCNN.

The first semi-supervised baseline (denoted as “DDCNN-semi”; “$M_8$”) is constructed by directly extending a supervised two-task DDCNN to a semi-supervised learning framework by computing consistency loss on saliency predictions for unlabeled RGB images with their associated pseudo depth maps. In other words, “DDCNN-semi” is equal to removing the consistency loss on the attention maps of our DS-Net. Furthermore, we experiment on only labeled data to explore the upper-bound performance of semi-supervised RGB-D saliency detection, where 3/10 labeled RGB-D data are split as the unlabeled RGB images. (denoted as “DDCNN-semi-ourSplit”; “$M_9$”). The last baseline (denoted as “DDCNN-semi-depthEst”; “$M_{10}$”) is designed to analyze the quality of estimated depth maps and the contribution of multi-task joint learning. “$M_{10}$” firstly utilizes an independent depth estimation model [55] to obtain paired RGB-D data for unlabeled RGB images, then trains the DS-Net in a semi-supervised manner. Among all the baseline networks, $M_1$ to $M_7$ are trained in a normal supervised manner with labeled RGB-D data. $M_8$ is also trained on labeled RGB-D data only but adopts fewer samples than supervised learning baselines. At last, $M_8$, $M_{10}$ and our DS-Net employ both labeled RGB-D data and unlabeled DUTS images.

Quantitative Comparisons. Table II lists the results of our network and ten baseline networks (i.e, $M_1$ to $M_{10}$) on six benchmark datasets. From the results, we have the following observations: (i) $M_7$ outperforms $M_1$, $M_2$ and $M_3$ in a large margin, showing that DAM, DGM and DIM have their contributions to the superior performance of our network. (ii) $M_7$ achieves larger $S_m$ and smaller $\text{MAE}$ scores than $M_4$, demonstrating that detecting additional depth maps enables DDCNN to better detect saliency maps from RGB-D image pairs. (iii) The superior results of $M_7$ over $M_5$ on two evaluation metrics indicates that DDCNN has a more accurate RGB-D saliency detection performance when computing a reconstruction loss on depth-aware RGB features and depth-dispersed RGB features since the reconstruction loss better separates these two kinds of RGB features. (iv) $M_7$ clearly outperforms $M_6$ in terms of $S_m$ and $\text{MAE}$ on six benchmark datasets, showing that DDCNN can better identify salient objects by using a pre-trained depth estimation branch for network initialization. (v) $M_8$ has better $S_m$ and $\text{MAE}$ results than $M_7$ on all six benchmark datasets. This shows that utilizing unlabeled RGB-D image pairs makes the RGB-D saliency detection of DDCNN more accurate, even though their depth maps of input are not the ground truths but are estimated by our DDCNN. (vi) Although $M_9$ fails to outperform both $M_7$ and $M_8$, it still wins a supervised method DMRA [10]. This result demonstrates that semi-supervised learning improves the capacity of supervised RGB-D saliency detectors and makes a main contribution to our final model. (vii) Compared to $M_8$, $M_{10}$ degrades performance, quantitatively verifying that our depth estimation branch can produce better depth maps than an independent model. In addition, $M_{10}$ has a competitive result with respect to $M_7$, indirectly proving that our DIM can merge depth and RGB features adaptively and discard the non-reliable depth pixels with the assistance of the attention mechanism. (viii) Our DS-Net can more accurately detect salient regions than $M_8$, as indicated by its superior results in two evaluation metrics on six benchmarks. This further shows that the consistency loss on attention maps contributes to the superior RGB-D saliency detection performance of our network.

Visual comparisons. Fig. 5 visually compares the saliency maps produced by our method and ten baseline networks. Among the results, our method can achieve the most similar saliency maps to the ground truth saliency, demonstrating the effectiveness of our DDCNN design and incorporating
unlabeled RGB images into RGB-D saliency detection.

C. Discussion

Evaluation of the depth maps. The depth estimation branch in our work is to align the labeled RGB-D data and unlabeled RGB images as inputs and does not limit the choice of the specific method. Although we only implement a small branch with a pretty simple architecture, it is still essential to evaluate the quality of depth maps generated by the branch. Specifically, we train the SG-Loss [56] and DBE [55] with the same setting parameters. Metrics summarized in Table V. From the results, our DS-Net significantly outperforms the baselines with the depth-dispelled features or depth-aware features only. It indicates that predicting RGB-D saliency with depth-dispelled features can recover the helpful information for saliency prediction is encoded into only $R_i^s$ and $R_i^d$, and corresponding depth features to utilize as much RGB information as possible. In addition, to further verify the role of $R_i^s$, we conduct a group of experiments where RGB-D saliency is predicted by fusing $R_i^s$ (or $R_i^d$) and depth features with our DAM (or DGM) only, which are denoted as “DDCNN-depthDispelledOnly” and “DDCNN-depthAwareOnly”, respectively. Table V shows the quantitative results. From the results, our DDCNN significantly outperforms the baselines with the depth-dispelled features or depth-aware features only. It indicates that predicting RGB-D saliency with depth-dispelled features can recover the RGB information as completely as possible and contributes to improving saliency detection accuracy.

Role of DIM in the semi-supervised learning setting. To further explore how the different components work under the semi-supervised learning setting, we construct three additional baselines denoted as “DS-Net-w/o-DAM”, “DS-Net-w/o-DGM”, and “DS-Net-w/o-DIM” by replacing DAM, DGM, and DIM of our DS-Net with a simple feature concatenation operation, respectively. Quantitative results are summarized in Table V. From the results, our DS-Net outperforms “DS-Net-w/o-DAM”, “DS-Net-w/o-DGM” and “DS-Net-w/o-DIM” by a substantial margin, demonstrating the necessity of DAM, DGM, and DIM in the semi-supervised learning setting. In addition, “DS-Net-w/o-DAM”, “DS-Net-w/o-DGM” and “DS-Net-w/o-DIM” win the corresponding supervised baselines (i.e., “DDCNN-w/o-DAM ($M_1$)”, “DDCNN-w/o-DGM ($M_2$)” and “DDCNN-w/o-DIM ($M_3$)”) in terms of six benchmark datasets, showing that the relatively weak baselines can also benefit from our semi-supervised framework.

Accuracy of the attention maps. Attention consistency is computed at multi-scale features from different convolutional layers. This contributes progressive and globally coherent guidance to semi-supervised learning, which ensures that each step of feature learning in our framework is relatively correct.
Table III: Quantitative results for depth estimation on NJU2K and NLPR datasets.

| Datasets | Our DDCNN | SG-Loss [2] | DBE [3] |
|----------|-----------|-------------|---------|
|          | MAE ↓ RMSE ↓ iMAE ↓ iRMSE ↓ | MAE ↓ RMSE ↓ iMAE ↓ iRMSE ↓ | MAE ↓ RMSE ↓ iMAE ↓ iRMSE ↓ |
| NJU2K    | 253.71    | 961.12 1.02 3.21 | 249.32 955.19 1.01 3.15 | 288.64 974.36 1.55 3.91 |
| NLPR     | 244.68    | 954.34 1.01 3.10 | 239.16 943.66 0.97 3.05 | 285.31 969.78 1.31 3.77 |

Table IV: Quantitative results of our method with different configurations of hyperparameters $\beta_1$ and $\beta_2$.

| Datasets | $\beta_1 = 1.0, \beta_2 = 1.0$ | $\beta_1 = 1.0, \beta_2 = 0.01$ | $\beta_1 = 0.01, \beta_2 = 10.0$ |
|----------|------------------------|---------------------|---------------------|
|          | $S_m \uparrow MAE \downarrow$ | $S_m \uparrow MAE \downarrow$ | $S_m \uparrow MAE \downarrow$ |
| NJU2k    | 0.825 0.096            | 0.818 0.066          | 0.950 0.965 0.966 0.024 |
| NLPR     | 0.800 0.111            | 0.788 0.078          | 0.936 0.933 0.961 0.021 |
| RGBD135  | 0.907 0.040            | 0.925 0.025          | 0.878 0.885 0.905 0.064 |
| LFSD     | 0.852 0.164            | 0.861 0.087          | 0.936 0.915 0.933 0.051 |
| SIP      | 0.860 0.084            | 0.894 0.066          | 0.886 0.915 0.933 0.051 |

Table V: Quantitative results of our method and baseline networks on NJU2K [40], NLPR [41], STERE [42], RGBD135 [43], LFSD [44], and SIP [3].

| Name             | NJU2K [40] | NLPR [41] | STERE [42] | RGBD135 [43] | LFSD [44] | SIP [3] |
|------------------|------------|------------|-------------|---------------|------------|---------|
|                  | $S_m \uparrow MAE \downarrow$ | $S_m \uparrow MAE \downarrow$ | $S_m \uparrow MAE \downarrow$ | $S_m \uparrow MAE \downarrow$ | $S_m \uparrow MAE \downarrow$ | $S_m \uparrow MAE \downarrow$ |
| “DDCNN-depthDispelledOnly” | 0.787 0.071 | 0.792 0.083 | 0.768 0.094 | 0.771 0.101 | 0.743 0.140 | 0.781 0.140 |
| “DDCNN-depthAwareOnly”   | 0.850 0.083 | 0.858 0.047 | 0.802 0.111 | 0.788 0.078 | 0.693 0.199 | 0.696 0.158 |
| our DDCNN          | 0.936 0.033 | 0.939 0.024 | 0.907 0.040 | 0.925 0.025 | 0.862 0.075 | 0.881 0.052 |
| “DDCNN-semi-linear” | 0.867 0.073 | 0.874 0.042 | 0.821 0.098 | 0.813 0.068 | 0.724 0.176 | 0.729 0.140 |
| our DDCNN-semi     | 0.941 0.025 | 0.942 0.020 | 0.911 0.038 | 0.934 0.024 | 0.869 0.071 | 0.883 0.051 |
| “DS-Net-w/o-DAM”   | 0.902 0.053 | 0.906 0.032 | 0.860 0.073 | 0.865 0.048 | 0.779 0.129 | 0.794 0.103 |
| “DS-Net-w/o-DGM”   | 0.909 0.049 | 0.913 0.030 | 0.868 0.068 | 0.875 0.044 | 0.801 0.120 | 0.807 0.095 |
| “DS-Net-w/o-DIM”   | 0.916 0.045 | 0.920 0.028 | 0.876 0.063 | 0.885 0.041 | 0.814 0.110 | 0.820 0.088 |
| “DS-Net-linear”    | 0.901 0.053 | 0.906 0.032 | 0.860 0.073 | 0.865 0.048 | 0.789 0.129 | 0.794 0.103 |
| our DS-Net         | 0.950 0.024 | 0.952 0.018 | 0.914 0.037 | 0.936 0.021 | 0.878 0.064 | 0.886 0.051 |

To further demonstrate the role of consistency loss on intermediate attention maps and evaluate their accuracy, we report our accuracy of the attention maps by using a linear evaluation protocol. Specifically, we extract the attention maps with four scales from the different layers of “DDCNN-semi(ln)” and our DS-Net, respectively, and train a logistic regression model to predict the final saliency with the resolution of 64 × 64, 32 × 32, 16 × 16 and 8 × 8. The training and testing split is consistent with the setting of the whole network. Table V quantitatively summarizes the results of linear evaluation with the resolution of 64 × 64. From the results, although “DS-Net-linear” degrades performance w.r.t. the normal DS-Net due to freezing the attention maps and most layers of the network, it outperforms the baseline “DDCNN-semi-linear” by a considerable margin. This demonstrates that the consistency loss on attention maps significantly improves the accuracy of the corresponding attention maps and contributes to the final saliency predictions.

Effects of the number of unlabeled images. In general, a model will gain a stronger generalization ability when fed into more samples. To further explore the relationship between the model performance and the number of unlabeled samples, we conduct five additional ablation experiments using our DS-Net on 1k, 2k, 4k, 8k, and 10k (10,553) unlabeled RGB images, respectively. All of the samples are randomly sampled from the DUTS dataset. Fig. 7 presents the effect of the number of unlabeled samples qualitatively. From the results, on the one hand, the model achieves higher and higher metric performance with the increase of the number of unlabeled samples, demonstrating that semi-supervised learning can benefit from more training data significantly. On the other hand, the performance gain decreases as more unlabeled data are fed into the model, indicating that the model has an upper-bound performance no matter how much data from the
same domain is employed. This indicates a research direction worth exploring, which is to improve further the generalization ability of the model by using out-of-domain data.

**Samples with multiple objects.** To study the upper-bound performance of our model and find the capability boundary of it, we delineate several samples with multiple objects. These samples are extremely challenging for existing state-of-the-art RGB-D saliency detectors, even for a human annotator. We carefully selected a few in Fig. 8. Apparently, our model can detect salient objects in these complicated samples accurately. This indicates that our model can handle samples with different levels of difficulty and achieve superior overall performance.

V. CONCLUSION

In this paper, we have presented a Dual-Semi RGB-D Salient Object Detection Network (DS-Net) by leveraging unlabeled RGB images to assist the RGB-D SOD task in a semi-supervised manner. A depth decoupling convolutional neural network (DDCNN) has been proposed to jointly estimate the pseudo depth maps for RGB images and predict saliency maps for paired RGB-D images. Our DDCNN disentangles the features from RGB images into depth-aware features and depth-dispersed features, which enables the network to identify the latent features specific to each modality. Experimental results on seven benchmark datasets have demonstrated the effectiveness of our method. We have also shown that the semi-supervised manner could further improve the performance of RGB-D SOD, even using an RGB image with the pseudo depth map.

ACKNOWLEDGEMENT

This work has been supported in part by the National Key Research and Development Program of China (2018AAA0101900), Zhejiang NSF (LR21F0200004), Chinese Knowledge Center of Engineering Science and Technology (CKCSET), Hikvision-Zhejiang University Joint Research Center, National Natural Science Foundation of China (Grant No. 61902275) and The Hong Kong Polytechnic University under Grant P0030419, Grant P0030929, and Grant P0035358.

REFERENCES

[1] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, “Review of visual saliency detection with comprehensive information,” IEEE Transactions on Circuits and Systems for Video Technology, 2018.

[2] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, “Saliency object detection in the deep learning era: An in-depth survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[3] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, “Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks,” IEEE Transactions on Neural Networks and Learning Systems, 2020.

[4] D. Feng, N. Barnes, S. You, and C. McCarthy, “Local background enclosure for RGB-D salient object detection,” in CVPR, 2016, pp. 2343–2350.

[5] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, “Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning,” IEEE Transactions on Image Processing, vol. 26, no. 9, pp. 4204–4216, 2017.

[6] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, “Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation,” IEEE Transactions on Image Processing, vol. 27, no. 2, pp. 568–579, 2017.

[7] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, “Going from RGB to RGBD saliency: A depth-guided transformation model,” IEEE Transactions on Cybernetics, 2019.

[8] H. Chen and Y. Li, “Progressively complementarity-aware fusion network for RGB-D salient object detection,” in CVPR, 2018, pp. 3051–3060.

[9] 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 CVPR, 2019, pp. 3927–3936.

[10] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, “Depth-induced multi-scale recurrent attention network for saliency detection,” in ICCV, 2019, pp. 7254–7263.

[11] 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 CVPR, 2020, pp. 3052–3062.

[12] 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 CVPR, 2020, pp. 8582–8591.

[13] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, “Select, supplement and focus for RGB-D saliency detection,” in CVPR, 2020, pp. 3472–3481.

[14] N. Liu, N. Zhang, and J. Han, “Learning selective self-mutual attention for RGB-D saliency detection,” in CVPR, 2020, pp. 13756–13765.

[15] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, “ASIF-Net: Attention Steered Interweave Fusion Network for RGB-D Salient Object Detection,” IEEE Transactions on Cybernetics, pp. 1–13, 2020.

[16] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, “Saliency object detection for RGB-D image by single stream recurrent convolution neural network,” Neurocomputing, vol. 363, pp. 46–57, 2019.

[17] P. Huang, C.-H. Shen, and H.-F. Hsiao, “RGBD salient object detection using spatially coherent deep learning framework,” in IEEE 23rd International Conference on Digital Signal Processing (DSP). IEEE, 2018, pp. 1–5.

[18] 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.

[19] N. Wang and X. Gong, “Adaptive fusion for RGB-D salient object detection,” IEEE Access, vol. 7, pp. 55277–55284, 2019.

[20] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, “Cross-modal weighting network for RGB-D salient object detection,” ECCV, 2020.

[21] Z. Chen, L. Zhu, L. Wan, S. Wang, W. Feng, and P.-A. Heng, “A multi-task mean teacher for semi-supervised shadow detection,” in CVPR, 2020, pp. 5611–5620.

[22] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, “Inf-Net: Autonomous COVID-19 Lung Infection Segmentation From CT Images,” IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2626–2637, aug 2020.

[23] S. Laine and T. Aila, “Temporal ensembling for semi-supervised learning,” in International Conference On Learning Representations, 2017.

[24] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in NIPS, 2017, pp. 1195–1204.

[25] 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.

[26] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “Cham: Convolutional block attention module,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[27] 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.

[28] H. Chen and Y. Li, “Contrast prior and fluid pyramid integration for RGBD salient object detection,” IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2825–2835, 2019.

[29] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, “Siamese network for rgb-d salient object detection and beyond,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[30] Y. Piao, L. Zhang, X. Zhang, and H. Lu, “Hierarchical dynamic filtering network for RGB-D salient object detection,” ECCV, 2020.

[31] S. Chen and Y. Fu, “Progressively guided alternate refinement network for RGB-D salient object detection,” ECCV, 2020.
[32] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, “A single stream network for robust and real-time RGB-D salient object detection,” ECCV, 2020.

[33] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, “RGB-D salient object detection with cross-modality modulation and selection,” ECCV, 2020.

[34] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu, “Cascade graph neural networks for RGB-D salient object detection,” in ECCV. Springer, 2020, pp. 346–364.

[35] M. Zhang, S. X. Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, “Asymmetric two-stream architecture for accurate RGB-D saliency detection,” in ECCV, 2020.

[36] Y. Wang, Y. Li, J. H. Elder, R. Wu, H. Lu, and L. Zhang, “Synergistic saliency and depth prediction for rgb-d saliency detection,” in Proceedings of the Asian Conference on Computer Vision, 2020.

[37] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, “Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1407–1417.

[38] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, “BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network,” in ECCV. Springer, 2020, pp. 275–292.

[39] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, “Accurate RGB-D salient object detection via collaborative learning,” ECCV, 2020.

[40] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, “Depth saliency based on anisotropic center-surround difference,” in IEEE international conference on image processing (ICIP). IEEE, 2014, pp. 1115–1119.

[41] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, “RGBD salient object detection: a benchmark and algorithms,” in ECCV. Springer, 2014, pp. 92–109.

[42] Y. Niu, Y. Geng, X. Li, and F. Liu, “Leveraging stereopsis for saliency analysis,” in CVPR. 2012, pp. 454–461.

[43] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, “Depth enhanced saliency detection method,” in Proceedings of international conference on internet multimedia computing and service. 2014, pp. 23–27.

[44] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, “Saliency detection on light field,” in CVPR, 2014, pp. 2806–2813.

[45] 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 CVPR, 2017, pp. 136–145.

[46] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in ICCV. 2017, pp. 4548–4557.

[47] C. Lang, J. Feng, S. Feng, J. Wang, and S. Yan, “Dual low-rank pursuit: Learning salient features for saliency detection,” IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 6, pp. 1190–1200, 2016.

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

[49] F. Perazzi, P. Krahenbühl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” in CVPR, 2012, pp. 733–740.

[50] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

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

[52] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, “High-resolution representations for labeling pixels and regions,” arXiv preprint arXiv:1904.04514, 2019.

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

[54] W. Liu, A. Rabinovich, and A. C. Berg, “ParseNet: Looking wider to see better,” arXiv preprint arXiv:1506.04579, 2015.

[55] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, “Single-image depth estimation based on fourier domain analysis,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 330–339.

[56] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, “Structure-guided ranking loss for single image depth prediction,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 611–620.

[57] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang, “Exploring disentangled feature representation beyond face identification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2080–2089.