Boosting RGB-D Saliency Detection by Leveraging Unlabeled RGB Images

Xiaoqiang Wang, Lei Zhu, Siliang Tang, Huazhu Fu, *Senior Member, IEEE*, Ping Li, *Member, IEEE*, Fei Wu, *Senior Member, IEEE*, Yi Yang, *Senior Member, IEEE*, and 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 semi-supervised DS-Net 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, attention consistency.**

I. INTRODUCTION

RGB-D salient object detection (SOD) has attracted a surge in interest recently [1]–[3]. Early RGB-D detectors [4]–[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]–[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 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
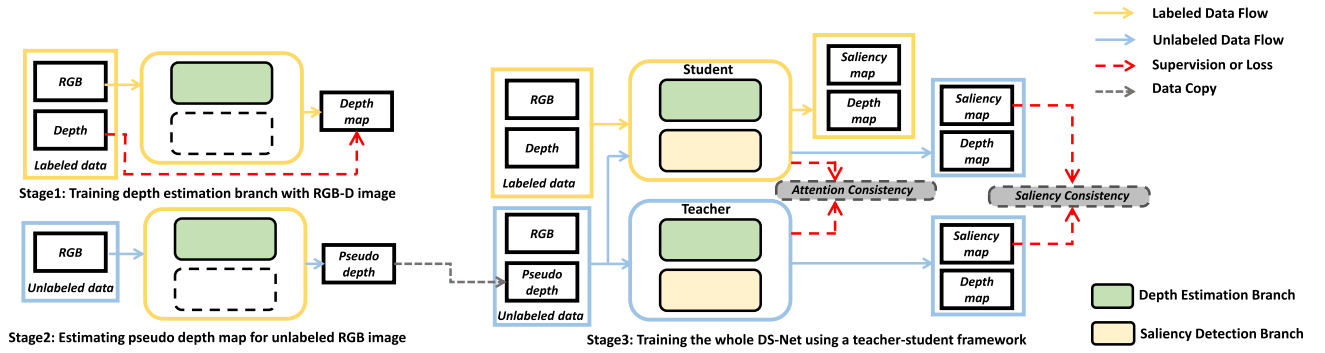
Fig. 1. Illustration of our DS-Net, which is trained in three main stages. 1) Training depth estimation branch based on labeled RGB-D data. 2) Estimating the pseudo depth maps for all unlabeled RGB images. 3) Learning the whole network in a semi-supervised manner based on labeled RGB-D data and unlabeled RGB images with pseudo 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 DDCNN 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 and Gong [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 and Li [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
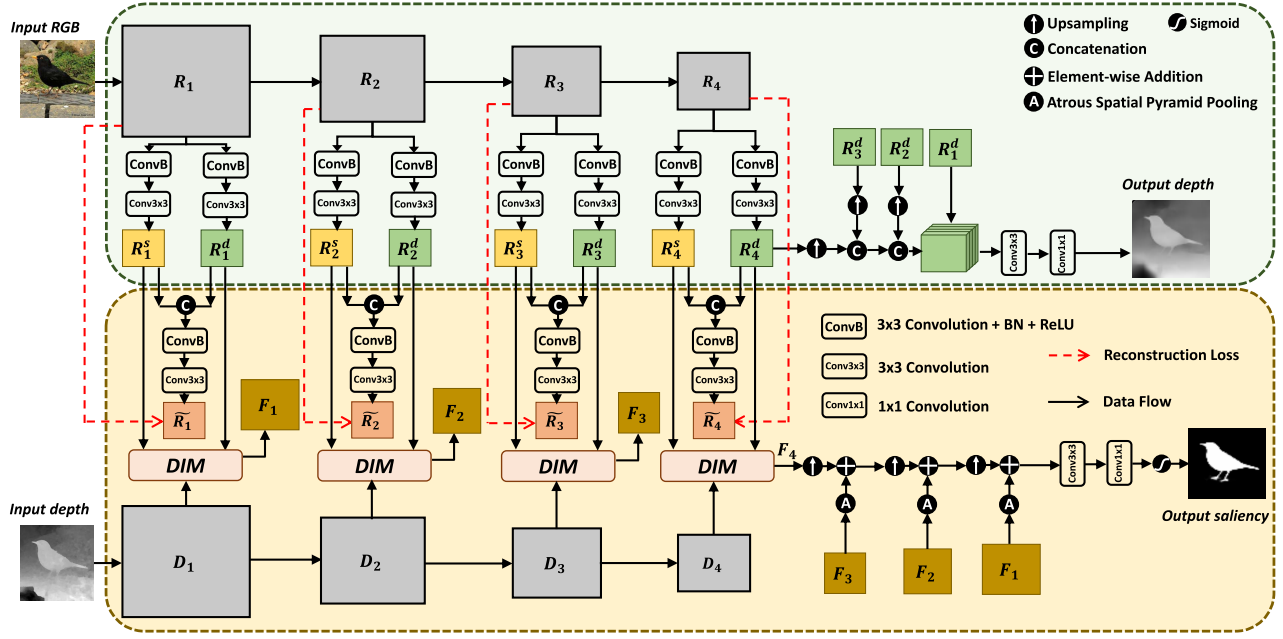
Fig. 2. Illustration of the proposed DDCNN, which includes a depth estimation branch (green block) and a saliency detection branch (yellow block). Stage 1 and stage 2 use the green block only, while stage 3 uses both green and yellow blocks. The depth estimation branch disentangles the image features R_i into depth-aware features R_i^d for estimating depth maps and depth-dispelled features R_i^s for predicting saliency. The saliency detection branch uses a depth-induced fusion module (DIM) to fuse depth-dispelled features R_i^s and depth features. Finally, the saliency map is predicted by fusing all features $\{F_i\}_{i=1}^4$ produced by DIMs.

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 with 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 saliency predictions for labeled RGB-D data and a consistency loss on intermediate attention maps and saliency predictions for unlabeled RGB images with their pseudo depth maps.

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). Given a pair of input RGB-D images, the RGB image is passed to an encoder to generate RGB features (*i.e.*, $\{R_i\}_{i=1}^4$), while the depth map is fed to another encoder to extract depth features (*i.e.*, $\{D_i\}_{i=1}^4$).

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-dispelled feature R_i^s for predicting saliency, using a convolutional block with "Conv$(3 \times 3) \rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Conv(3×3)". The depth-aware features R_i^d are then up-sampled to the same spatial resolution of R_1^d and concatenated together to predict a depth map by applying a "Conv$(3 \times 3) \rightarrow$ 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 $\mathcal{L}_i^r$ to regularize the decoupling process:

$$\mathcal{L}_i^r = L_{MSE}(\tilde{R}_i, R_i), \text{ with } \tilde{R}_i = CB\left(Cat\left(R_i^d, R_i^s\right)\right), \quad (1)$$
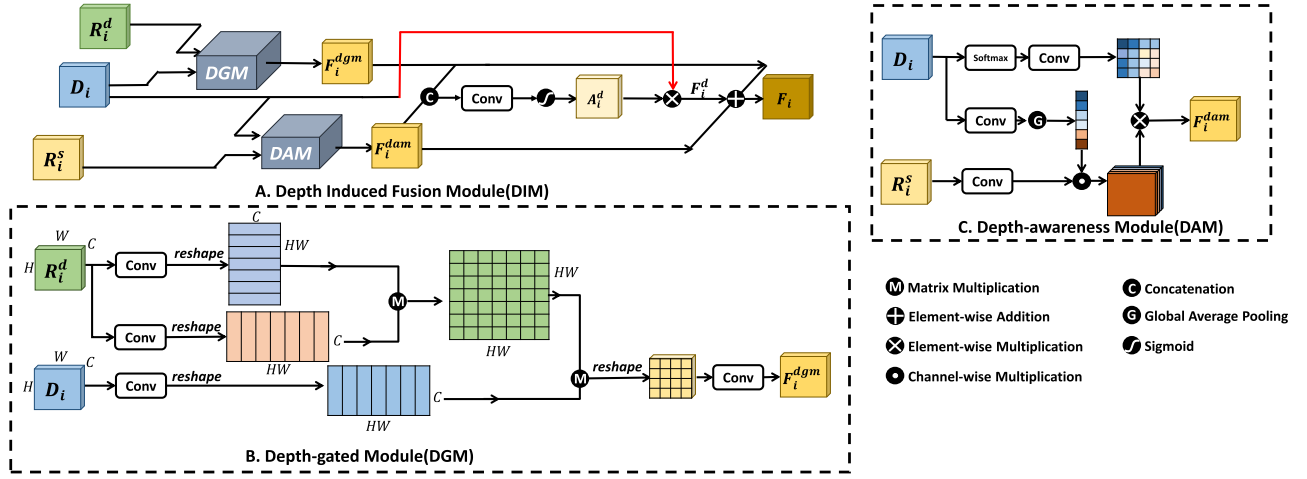
Fig. 3. The illustration of DIM, which contains the Depth-awareness Module (DAM) and Depth-gated Module (DGM).

where $CB(\cdot)$ is a convolutional block with "Conv$(3 \times 3) \rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ 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-dispelled 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. When merging features (F_i and 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 1×1 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}.

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-dispelled 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 3×3 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:

$$F_i = F_i^{dam} + F_i^{dgm} + F_i^d, \qquad (2)$$

2) 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}:

$$F_i^{dam} = S_{att}(D_i) \otimes (C_{att}(D_i) \odot R_i^s), \qquad (3)$$

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

3) 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 3×3 convolutional layer on R_i^d, reshape the resultant features as an intermediate $C \times HW$ feature map, apply another 3×3 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 3×3 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 3×3 convolution to obtain the features F_i^{dgm} of DGM.

B. Training Strategies

1) 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 and G_d), the supervised loss (denoted as $\mathcal{L}^s$) for a labeled image (x) is computed as the summation of the saliency detection loss and depth estimation loss:

$$\mathcal{L}^s(x) = L_{BCE}(P_s, G_s) + \alpha \, L_{MSE}(P_d, G_d), \quad (4)$$

where P_s 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.

2) Consistency Loss for Unlabeled Data: As a well-known semi-supervised manner, the teacher-student approach ensembles parameters of the network at different training processes 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 DDCNN to predict a depth map and take this prediction as a pseudo label of the depth map for the unlabeled RGB data. After that, we pass the unlabeled RGB image 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 force the predicted saliency map and the four learned attention maps of DIMs to be consistent for the student and teacher networks. A consistency loss $\mathcal{L}^c$ on unlabeled data y is defined as:

$$\mathcal{L}^c(y) = L_{MSE}(S_s, T_s) + \gamma \sum_{l=1}^{4} L_{MSE}(S_A^l, T_A^l), \quad (5)$$

where S_s and T_s denote two detected saliency maps of the student and the teacher networks. S_A^l and T_A^l represent the learned attention map of DIM at the l-th CNN layer from the student and the teacher networks. We empirically set $\gamma = 0.1$ in this work.

3) Overall Loss for the Student Network: The total loss of our network is computed as:

$$\mathcal{L}_{total} = \frac{1}{N_1} \sum_{i=1}^{N_1} \left(\mathcal{L}^s(x_i) + \beta_1 \sum_{l=1}^{4} \mathcal{L}_l^r(x_i) \right)$$
$$+ \lambda \frac{1}{N_2} \sum_{j=1}^{N_2} \left(\mathcal{L}^c(y_j) + \beta_2 \sum_{l=1}^{4} \mathcal{L}_l^r(y_j) \right), \quad (6)$$

where N_1 and N_2 are the number of labeled RGB-D image pairs and unlabeled RGB images in our training set.

$\mathcal{L}^s(x_i)$ denotes the supervised loss (Eq. (4)) for the i-th labeled image x_i while $\mathcal{L}^c(y_j)$ is the consistency loss (Eq. (5)) for the j-th unlabeled image y_j. $\mathcal{L}_l^r(x_i)$ and $\mathcal{L}_l^r(y_j)$ 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 λ balances the loss between the labeled and unlabeled data. Following [21], we use a time dependent Gaussian warm-up function to update λ: $\lambda(t) = \lambda_{max} e^{(-5(1-t/t_{max})^2)}$, where t denotes the current training iteration and t_{max} is the maximum training iteration. In our experiments, we empirically set $\lambda_{max} = 1.0$.

Following existing self-ensembling frameworks [24], we minimize the total loss $\mathcal{L}_{total}$ 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.

4) Training Configurations: We use the same HRNet architecture as the backbone for both the RGB image and the depth map. Given a 256×256 input image, the HRNet outputs multi-scale features at four convolutional blocks (layer 8, 27, 63, and 104) with a resolution of 64×64, 32×32, 16×16, and 8×8, 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

A. 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) LFSD [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 [3], [9], [11] 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.

B. 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_β^{max}) [47], E-measure

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	LBE[4]	DF[27]	CTMF[18]	PCF[8]	TANet[28]	CPFP[9]	SSF[13]	UCNet[12]	JLDCF[11]	JLDCF[J][29]	HDF-Net[30]	PGA-Net[31]	DANet[32]	cmMS[33]	Cas-Gnn[34]	CMWNet[20]	VGG-16 DS-Net	DMRA[10]	ATSA[35]	SSDP[36]	DSA²F[37]	VGG-19 DS-Net	D³Net[3]	BBS-Net[38]	CoNet[39]	ResNet-50 DS-Net	DDCNN-semi ourSplit	HRNet DDCNN	HRNet DS-Net
NJU2K	$S_m\uparrow$	0.695	0.763	0.849	0.877	0.878	0.879	0.899	0.897	0.903	0.911	0.908	0.906	0.901	0.904	0.911	0.903	**0.945**	0.886	0.899	0.878	0.903	**0.944**	0.895	0.921	0.894	**0.946**	0.922	0.936	**0.950**
	$F_\beta^{max}\uparrow$	0.748	0.804	0.845	0.872	0.874	0.877	0.886	0.886	0.903	0.913	0.922	0.883	0.893	0.914	0.903	0.902	**0.961**	0.886	0.910	0.852	0.901	**0.964**	0.889	0.920	0.872	**0.964**	0.943	0.958	**0.965**
	$E_\phi^{max}\uparrow$	0.803	0.864	0.913	0.924	0.925	0.926	-	0.930	0.944	0.948	0.932	0.914	0.921	-	0.936	0.933	**0.962**	0.927	0.922	0.909	0.923	**0.964**	0.932	0.949	0.912	**0.964**	0.944	0.954	**0.966**
	$MAE\downarrow$	0.153	0.141	0.085	0.059	0.060	0.053	0.043	0.043	0.043	0.040	0.038	0.045	0.040	0.044	0.035	0.046	**0.026**	0.051	0.045	0.055	0.039	**0.026**	0.051	0.035	0.047	**0.025**	0.037	0.033	**0.024**
NLPR	$S_m\uparrow$	0.762	0.802	0.860	0.874	0.886	0.888	0.914	0.920	0.925	0.926	0.923	0.918	0.907	0.900	0.919	0.917	**0.949**	0.899	0.915	0.875	0.918	**0.950**	0.905	0.930	0.907	**0.951**	0.922	0.939	**0.952**
	$F_\beta^{max}\uparrow$	0.745	0.778	0.825	0.841	0.863	0.867	0.875	0.891	0.916	0.917	0.927	0.871	0.876	0.914	0.904	0.903	**0.948**	0.879	0.916	0.809	0.897	**0.948**	0.885	0.918	0.848	**0.951**	0.919	0.938	**0.953**
	$E_\phi^{max}\uparrow$	0.855	0.880	0.929	0.925	0.941	0.932	-	0.951	0.962	0.964	0.957	0.948	0.945	-	0.952	0.951	**0.965**	0.947	0.949	0.915	0.950	**0.966**	0.946	0.961	0.936	**0.968**	0.942	0.958	**0.970**
	$MAE\downarrow$	0.081	0.085	0.056	0.044	0.041	0.036	0.026	0.025	0.023	0.023	0.023	0.028	0.028	0.273	0.025	0.029	**0.020**	0.031	0.028	0.044	0.024	**0.020**	0.034	0.023	0.031	**0.018**	0.032	0.024	**0.018**
STERE	$S_m\uparrow$	0.660	0.757	0.848	0.875	0.871	0.879	0.893	0.903	0.905	0.907	0.900	0.897	-	0.889	0.899	0.905	**0.909**	0.886	0.903	0.893	-	**0.910**	0.891	0.908	0.908	**0.910**	0.892	0.907	**0.914**
	$F_\beta^{max}\uparrow$	0.633	0.757	0.831	0.860	0.861	0.874	0.880	0.884	0.901	0.907	0.910	0.884	-	0.908	0.901	0.901	**0.912**	0.886	0.872	0.878	-	**0.914**	0.881	0.903	0.885	**0.914**	0.896	0.910	**0.915**
	$E_\phi^{max}\uparrow$	0.787	0.847	0.912	0.925	0.923	0.925	-	0.935	**0.936**	0.935	0.931	0.921	-	-	0.930	0.934	0.935	**0.937**	0.914	0.936	-	**0.937**	0.930	0.932	0.923	**0.937**	0.911	0.937	**0.947**
	$MAE\downarrow$	0.250	0.141	0.086	0.064	0.060	0.051	0.044	**0.039**	0.042	0.039	0.041	0.039	-	0.042	**0.039**	0.043	0.039	0.047	0.044	0.045	-	**0.039**	0.054	0.041	0.041	**0.039**	0.041	0.040	**0.037**
RGBD135	$S_m\uparrow$	0.703	0.752	0.863	0.842	0.858	0.872	0.905	0.934	0.929	0.931	0.926	0.894	0.907	-	0.905	0.934	**0.934**	0.900	0.924	0.890	0.904	**0.935**	0.904	0.933	0.910	**0.934**	0.902	0.925	**0.936**
	$F_\beta^{max}\uparrow$	0.788	0.766	0.844	0.804	0.827	0.846	0.876	0.919	0.919	0.929	**0.932**	0.870	0.885	-	0.906	0.930	0.932	0.888	0.928	0.864	0.898	**0.932**	0.885	0.927	0.861	**0.932**	0.908	0.929	**0.933**
	$E_\phi^{max}\uparrow$	0.890	0.870	0.932	0.893	0.910	0.923	-	0.967	0.968	0.955	**0.971**	0.935	0.952	-	0.947	0.969	0.959	0.943	**0.968**	0.927	0.933	0.960	0.946	**0.966**	0.945	0.960	0.913	0.955	0.961
	$MAE\downarrow$	0.208	0.093	0.055	0.049	0.041	0.038	0.025	**0.019**	0.022	0.022	0.021	0.032	0.024	-	0.028	0.022	0.021	0.030	0.023	0.031	0.036	**0.021**	0.030	0.021	0.027	**0.021**	0.032	0.025	0.021
LFSD	$S_m\uparrow$	0.729	0.783	0.788	0.786	0.794	0.820	0.859	0.854	0.854	0.863	0.854	0.855	-	0.860	0.849	0.856	**0.866**	0.839	0.833	0.830	0.920	**0.869**	0.824	0.854	0.862	**0.872**	0.842	0.862	**0.878**
	$F_\beta^{max}\uparrow$	0.722	0.813	0.787	0.775	0.792	0.821	0.867	0.855	0.862	0.862	0.883	0.862	-	0.883	0.864	0.883	**0.884**	0.852	0.830	0.823	0.896	**0.885**	0.815	0.858	0.848	**0.884**	0.863	0.883	**0.885**
	$E_\phi^{max}\uparrow$	0.797	0.857	0.857	0.827	0.840	0.864	-	0.901	0.893	0.900	0.891	0.900	-	-	0.877	0.902	0.902	0.893	0.869	0.879	0.962	**0.903**	0.856	0.901	0.897	**0.905**	0.874	0.902	**0.905**
	$MAE\downarrow$	0.214	0.146	0.127	0.119	0.118	0.095	0.086	0.086	0.078	0.079	**0.076**	0.086	-	0.082	0.083	0.086	**0.076**	0.083	0.093	0.090	0.021	**0.073**	0.106	0.072	0.071	**0.069**	0.082	0.075	**0.064**
SIP	$S_m\uparrow$	0.727	0.653	0.716	0.842	0.835	0.850	-	0.875	0.879	0.882	**0.886**	0.875	0.875	-	0.867	0.881	0.806	-	0.880	0.882	**0.883**	0.864	0.879	0.858	**0.883**	0.856	0.881	**0.886**
	$F_\beta^{max}\uparrow$	0.751	0.657	0.694	0.838	0.830	0.851	-	0.867	0.885	0.900	0.901	0.892	0.848	-	0.874	0.909	0.821	-	0.856	0.882	**0.909**	0.862	0.883	0.842	**0.912**	0.877	0.902	**0.915**
	$E_\phi^{max}\uparrow$	0.853	0.759	0.829	0.901	0.895	0.903	-	0.914	0.923	0.919	0.922	0.915	0.908	-	0.913	0.924	0.875	-	0.922	0.923	**0.925**	0.910	0.922	0.909	**0.927**	0.886	0.923	**0.933**
	$MAE\downarrow$	0.200	0.185	0.139	0.071	0.075	0.064	-	0.051	0.051	0.058	0.057	0.054	0.059	-	0.062	0.053	0.085	-	0.052	0.054	**0.055**	0.063	0.055	0.063	**0.051**	0.070	0.052	**0.051**
DUTD	$S_m\uparrow$	-	0.695	0.499	-	0.526	0.736	0.791	0.801	0.808	-	0.818	-	0.899	0.903	-	-	**0.913**	0.702	0.889	0.885	0.921	**0.916**	0.831	-	0.898	**0.917**	0.912	0.923	**0.928**
	$F_\beta^{max}\uparrow$	-	0.692	0.411	-	0.458	0.740	0.767	0.771	0.790	-	0.898	-	0.918	0.901	-	-	**0.919**	0.659	0.795	0.878	0.926	**0.920**	0.823	-	0.903	**0.924**	0.915	0.929	**0.932**
	$E_\phi^{max}\uparrow$	-	0.800	0.654	-	0.709	0.823	0.859	0.856	0.861	-	0.859	-	0.937	0.937	-	-	**0.938**	0.796	0.933	0.935	0.950	**0.940**	0.899	-	0.931	**0.940**	0.927	0.939	**0.942**
	$MAE\downarrow$	-	0.220	0.243	-	0.201	0.144	0.113	0.100	0.093	-	0.076	-	0.043	0.043	-	-	**0.036**	0.122	0.048	0.057	0.030	**0.035**	0.097	-	0.045	**0.036**	0.045	0.042	0.035

TABLE II

QUANTITATIVE RESULTS OF OUR METHOD AND BASELINE NETWORKS FOR THE ABLATION EXPERIMENTS ON NJU2K [40], NLPR [41], STERE [42], RGBD135 [43], LFSD [44], AND SIP [3]. "SL" DENOTES SUPERVISED LEARNING, WHILE "SSL" REPRESENTS SEMI-SUPERVISED LEARNING

Name	Networks	SL	SSL	NJU2K [40] $S_m\uparrow$	$MAE\downarrow$	NLPR [41] $S_m\uparrow$	$MAE\downarrow$	STERE [42] $S_m\uparrow$	$MAE\downarrow$	RGBD135 [43] $S_m\uparrow$	$MAE\downarrow$	LFSD [44] $S_m\uparrow$	$MAE\downarrow$	SIP [3] $S_m\uparrow$	$MAE\downarrow$
M_1	"DDCNN-w/o-DAM"	✓		0.881	0.065	0.887	0.038	0.837	0.088	0.834	0.060	0.750	0.157	0.755	0.125
M_2	"DDCNN-w/o-DGM"	✓		0.893	0.054	0.888	0.036	0.848	0.071	0.859	0.056	0.765	0.148	0.762	0.115
M_3	"DDCNN-w/o-DIM"	✓		0.835	0.089	0.878	0.042	0.659	0.176	0.811	0.062	0.652	0.201	0.672	0.172
M_4	"DDCNN-w/o-depth"	✓		0.855	0.079	0.888	0.037	0.698	0.162	0.829	0.056	0.669	0.194	0.685	0.164
M_5	"DDCNN-w/o-reconstr-loss"	✓		0.894	0.051	0.908	0.038	0.835	0.081	0.899	0.042	0.744	0.144	0.766	0.110
M_6	"DDCNN-w/o-pretrain"	✓		0.934	0.033	0.939	0.023	0.902	0.045	0.919	0.027	0.855	0.079	0.866	0.061
M_7	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
M_8	"DDCNN-semi"		✓	0.941	0.025	0.942	0.020	0.911	0.038	0.934	0.024	0.869	0.071	0.883	0.052
M_9	"DDCNN-semi-ourSplit"		✓	0.922	0.037	0.922	0.032	0.892	0.041	0.902	0.032	0.842	0.082	0.856	0.070
M_{10}	"DDCNN-semi-depthEst"		✓	0.938	0.031	0.941	0.022	0.906	0.039	0.926	0.025	0.866	0.074	0.881	0.053
	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**

(E_ϕ^{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_β^{max}, a larger E_ϕ^{max}, and a smaller MAE.

C. 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.

D. 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], D³Net [3], SSF [13], UCNet [12], JLDCF [11], JLDCF[J] [29], HDF-Net [30], ATSA [35], SSDP [36], DSA²F [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_β^{max}, E_ϕ^{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

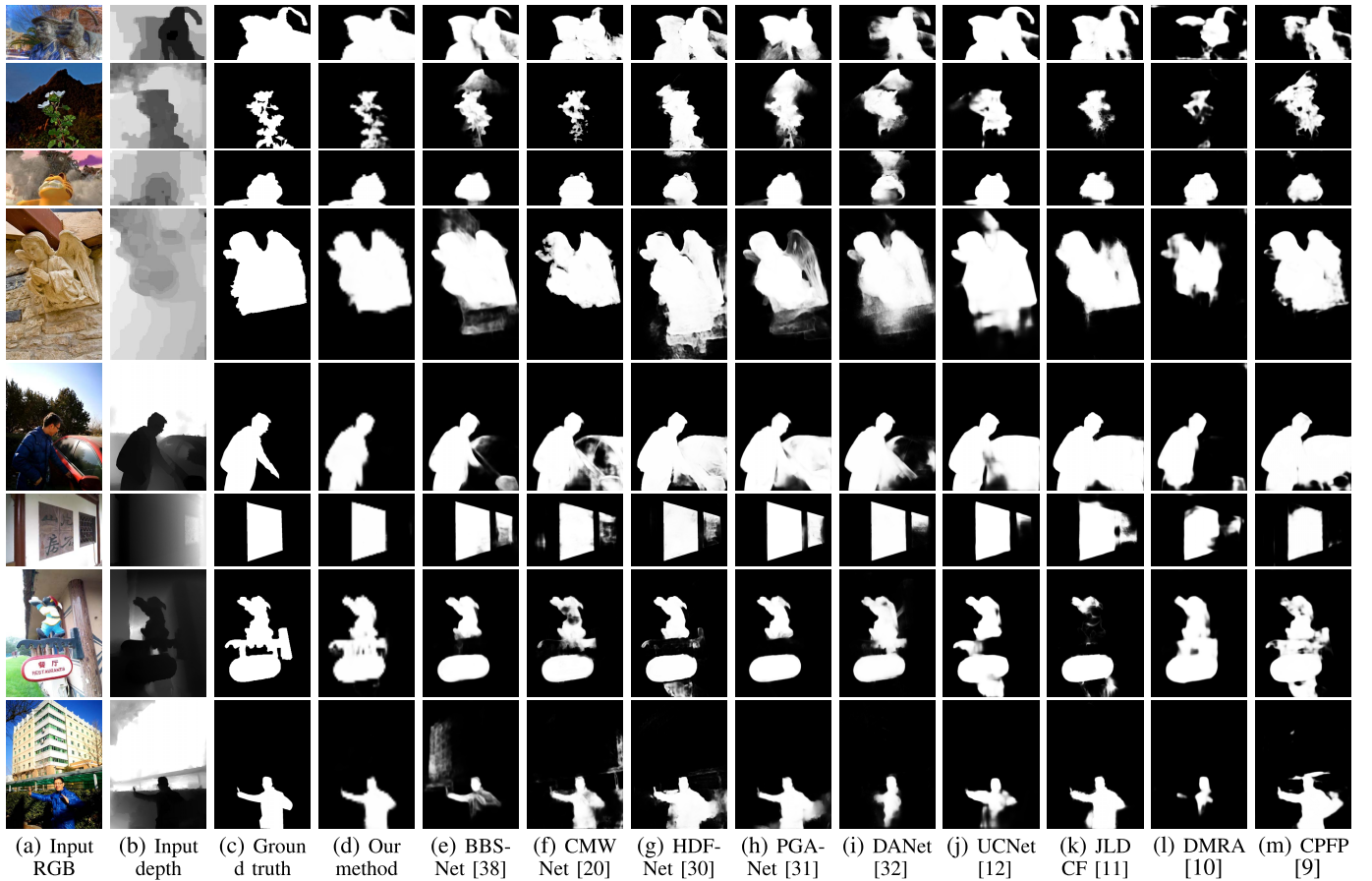| (a) Input RGB | (b) Input depth | (c) Ground truth | (d) Our method | (e) BBS-Net [38] | (f) CMW Net [20] | (g) HDF-Net [30] | (h) PGA-Net [31] | (i) DANet [32] | (j) UCNet [12] | (k) JLD CF [11] | (l) DMRA [10] | (m) CPFP [9] |

Fig. 4. Visual comparison of saliency map results produced by different methods. As can be seen, our network produces more accurate saliency maps than compared methods.

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_β^{max} improvement of 1.63%, a E_ϕ^{max} improvement of 0.66%, and a 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.

E. Ablation Analysis

1) 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-reconstr-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 DDCNN.

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

(a) RGB　(b) depth　(c) GT　(d) ours　(e) M_{10}　(f) M_9　(g) M_8　(h) M_7　(i) M_6　(j) M_5　(k) M_4　(l) M_3　(m) M_2　(n) M_1
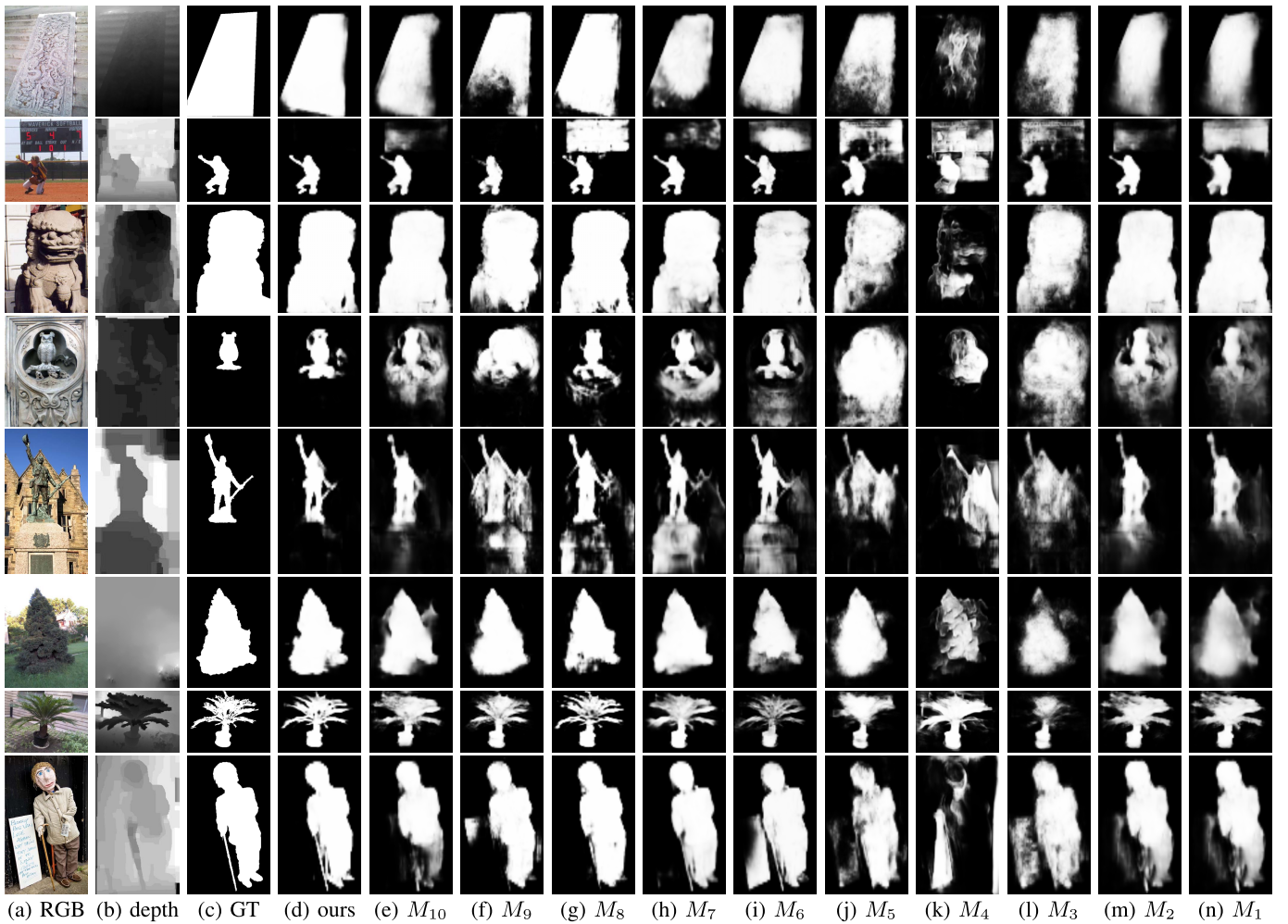
Fig. 5.　Visual comparison of saliency map results produced by different methods. (a) Input RGB image from benchmark datasets; (b) Input depth image; (c) Ground truths (denoted as 'GT'); (d)–(n) Saliency maps predicted by our method and ten constructed baseline networks (i.e., M_{10} to M_1). Please refer to Section IV-E for the explanation of M_1 to M_{10}.

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_9 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.

2) 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 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-dispelled* 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 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 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,

TABLE III
QUANTITATIVE RESULTS FOR DEPTH ESTIMATION ON NJU2K AND NLPR DATASETS

Datasets	Our DDCNN				SG-Loss [2]				DBE [3]			
	$MAE\downarrow$	$RMSE\downarrow$	$iMAE\downarrow$	$iRMSE\downarrow$	$MAE\downarrow$	$RMSE\downarrow$	$iMAE\downarrow$	$iRMSE\downarrow$	$MAE\downarrow$	$RMSE\downarrow$	$iMAE\downarrow$	$iRMSE\downarrow$
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

(a) RGB   (b) depth   (c) ours   (d) sg-loss   (e) DBE
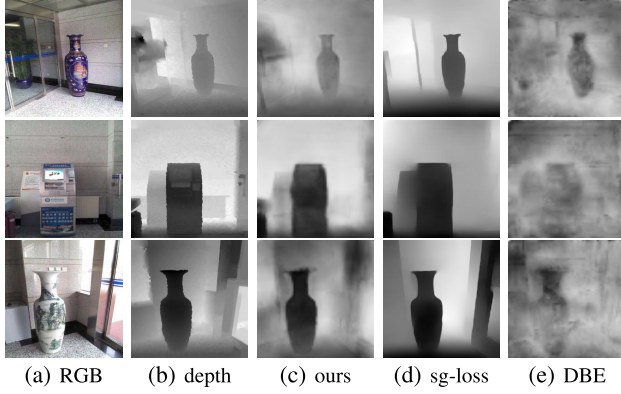
Fig. 6.   Qualitative results of our depth estimation branch.

Fig. 7.   Illustration of performance (in S_m and E_ϕ^{max}) of our model with different hyperparameter configurations and numbers of unlabeled images.

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.

3) Visual Comparisons: Fig. 5 visually compares the saliency maps predicted 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.

F. Discussion

1) 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 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 as the depth estimation branch of our DDCNN. We adopt four widely used metrics, including Mean Average Error (MAE), Root Mean Square Error (RMSE), and their inverses. Although only the depth maps for unlabeled DUTS images are fed into our DS-Net, they have no available paired ground truth depth maps for evaluation. Thus,
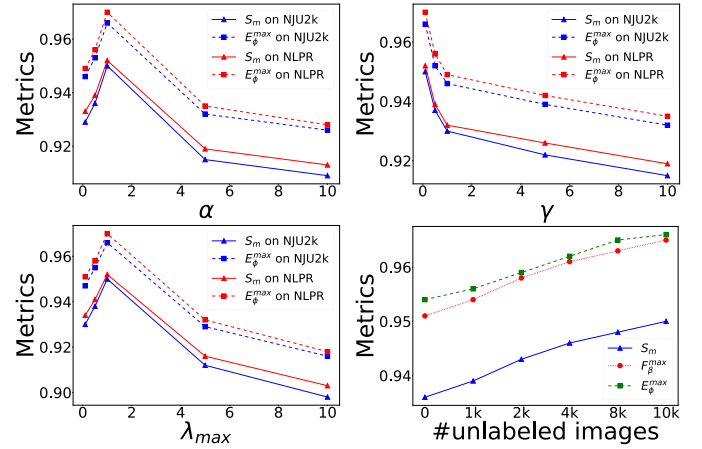
we report the results on the unseen testing set of NJU2K and NLPR benchmarks. Quantitative and qualitative results are summarized in Table III and Fig. 6. From the results, although our depth estimation branch is devised with a pretty simple decoder, it still outperforms the DBE in a significant margin and achieves comparable results against the more recent depth estimation model (*i.e.*SG-Loss). These results show that our feature disentangling can extract depth-related features correctly and ensure the quality of pseudo depth maps for unlabeled RGB images. What is more, the powerful feature encoder also regularizes the RGB-D saliency detection to make up for the lack of perfect depth quality.

2) Effects of the Hyperparameters: We conduct a group of experiments to select the best configurations for hyperparameters α in Eq. 4, γ in Eq. 5 and β_1, β_2 and λ_{max} in Eq. 6. Table IV and Fig. 7 summarize the results quantitatively and qualitatively. From the results, the configuration with $\alpha = 1.0$, $\gamma = 0.1$, $\beta_1 = 0.01$ and $\beta_2 = 1.0$ and $\lambda_{max} = 1.0$ achieves the best results, which is consistent with our empirical setting obtained after a few warm-up epochs. The reason behind this is that the best configuration may be the one maintaining the best balance for the order of magnitudes during the whole training process.

3) Role of the Depth-Dispelled Features: In this paper, R_i^s is designed to predict the RGB-D saliency. However, because of the possible non-equivalent mapping of RGB features disentangling [57]–[59], we cannot guarantee that all the helpful information for saliency prediction is encoded into only R_i^s. Consequently, while predicting the final RGB-D saliency, we merge R_i^s, R_i^d, and corresponding depth features to utilize as much RGB information as possible. In addition,

TABLE IV

QUANTITATIVE RESULTS OF OUR METHOD WITH DIFFERENT CONFIGURATIONS OF HYPERPARAMETERS β_1 AND β_2

Datasets	$\beta_1=1.0, \beta_2=1.0$				$\beta_1=1.0, \beta_2=0.01$				$\beta_1=0.01, \beta_2=1.0$				$\beta_1=0.01, \beta_2=10.0$			
	$S_m\uparrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$MAE\downarrow$	$S_m\uparrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$MAE\downarrow$
NJU2k	0.929	0.943	0.946	0.036	0.936	0.951	0.952	0.032	**0.950**	**0.965**	**0.966**	**0.024**	0.943	0.958	0.959	0.028
NLPR	0.933	0.932	0.949	0.024	0.939	0.938	0.956	0.022	**0.952**	**0.953**	**0.970**	**0.018**	0.945	0.946	0.963	0.020
STERE	0.891	0.892	0.920	0.052	0.898	0.900	0.929	0.047	**0.914**	**0.915**	**0.947**	**0.037**	0.906	0.907	0.938	0.042
RGBD135	0.905	0.904	0.930	0.033	0.915	0.914	0.940	0.029	**0.936**	**0.933**	**0.961**	**0.021**	0.926	0.923	0.951	0.025
LFSD	0.840	0.849	0.864	0.092	0.852	0.861	0.878	0.083	**0.878**	**0.885**	**0.905**	**0.064**	0.865	0.873	0.891	0.073
SIP	0.846	0.869	0.889	0.073	0.860	0.884	0.904	0.066	**0.886**	**0.915**	**0.933**	**0.051**	0.873	0.899	0.919	0.058

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"	0.870	0.071	0.878	0.041	0.825	0.096	0.818	0.066	0.731	0.171	0.735	0.136
"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	0.789	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 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.

4) 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.

5) 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. 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(M_8)" 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.

6) 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
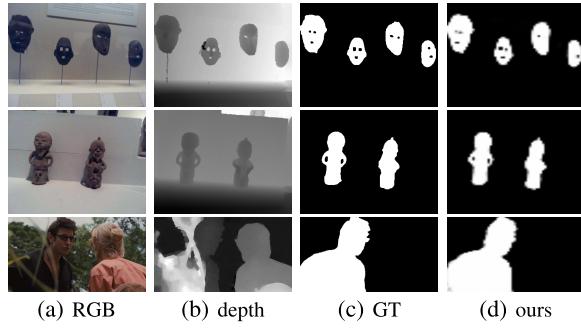
(a) RGB (b) depth (c) GT (d) ours

Fig. 8. Samples with multiple objects. (a) Input RGB image from benchmark datasets; (b) Input depth image; (c) Ground truths (denoted as 'GT'); and (d) Our results.

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.

7) 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-dispelled 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.

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 Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[2] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 12, 2021, doi: 10.1109/TPAMI.2021.3051099.

[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 Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[4] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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 Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, Sep. 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 Trans. Image Process.*, vol. 27, no. 2, pp. 568–579, Feb. 2018.

[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 Trans. Cybern.*, vol. 50, no. 8, pp. 3627–3639, Aug. 2020.

[8] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3052–3062.

[12] J. Zhang *et al.*, "UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3472–3481.

[14] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13756–13765.

[15] C. Li *et al.*, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 2021.

[16] 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, Oct. 2019.

[17] P. Huang, C.-H. Shen, and H.-F. Hsiao, "RGBD salient object detection using spatially coherent deep learning framework," in *Proc. IEEE 23rd Int. Conf. Digit. Signal Process. (DSP)*, Nov. 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 Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.

[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," in *Proc. ECCV*, 2020, pp. 665–681.

[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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5611–5620.

[22] D. Fan *et al.*, "Inf-Net: Automatic COVID-19 lung infection segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, Aug. 2020.

[23] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, *arXiv:1610.02242*.

[24] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[26] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (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 Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.

[28] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 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 Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 16, 2021, doi: 10.1109/TPAMI.2021.3073689.

[30] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. ECCV*, 2020, pp. 235–252.

[31] S. Chen and Y. Fu, "Progressively guided alternate refinement network for RGB-D salient object detection," in *Proc. ECCV*, 2020, pp. 520–538.

[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," in *Proc. ECCV*, 2020, pp. 646–662.

[33] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "RGB-D salient object detection with cross-modality modulation and selection," in *Proc. ECCV*, 2020, pp. 225–241.

[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 *Proc. ECCV*. Cham, Switzerland: 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 *Proc. ECCV*, 2020, pp. 374–390.

[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 *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–17.

[37] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep RGB-D saliency detection with depth-sensitive attention and automatic multimodal fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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 *Proc. ECCV*. Cham, Switzerland: 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," in *Proc. ECCV*, 2020, pp. 52–69.

[40] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 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 *Proc. ECCV*. Cham, Switzerland: Springer, 2014, pp. 92–109.

[42] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.

[43] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2014, pp. 23–27.

[44] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[45] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 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 Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1190–1200, Jun. 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 *Proc. IJCAI*, 2018, pp. 698–704.

[49] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

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

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[52] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

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

[55] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on Fourier domain analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2080–2089.

[58] J.-T. Hsieh, B. Liu, D.-A. Huang, L. Fei-Fei, and J. C. Niebles, "Learning to decompose and disentangle representations for video prediction," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 515–524.

[59] J. Ren, M. Li, Z. Liu, and Q. Zhang, "Interpreting and disentangling feature components of various complexity from DNNs," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8971–8981.

Xiaoqiang Wang received the B.Eng. degree from Zhejiang University, Hangzhou, China, where he is currently pursuing the master's degree with the College of Computer Science and Technology. His research interests are computer vision and machine learning, with a particular focus on multimodal data analysis.

Lei Zhu received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong. He is currently working as an Assistant Professor with the ROAS Thrust, HKUST (GZ), and also an affiliated Assistant Professor in ECE with HKUST. Before that, he was a Postdoctoral Researcher at DAMTP, University of Cambridge. His research interests include computer graphics, computer vision, medical image processing, and deep learning.
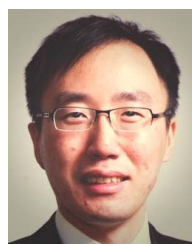
Siliang Tang received the Ph.D. degree from the National University of Ireland, Maynooth, Ireland, in 2012. He is currently an Associate Professor with the College of Computer Science, Zhejiang University. His research interests include information extraction, knowledge base construction, and multimodal data analysis.

Huazhu Fu (Senior Member, IEEE) received the Ph.D. degree from Tianjin University in 2013. He is currently a Senior Scientist at IHPC, A*STAR. Previously, he was a Research Fellow (2013–2015) with Nanyang Technological University (NTU), Singapore; a Research Scientist (2015–2018) with the Institute for Infocomm Research (I2R), A*STAR, Singapore; and a Senior Scientist (2018–2021) with the Inception Institute of Artificial Intelligence (IIAI), United Arab Emirates. His research interests include computer vision, machine learning, and AI in healthcare.
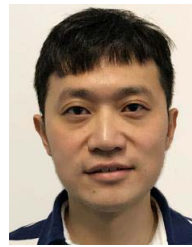
Yi Yang (Senior Member, IEEE) receiving the Ph.D. degree in computer science from Zhejiang University in 2010. He is currently a Professor with the Faculty of Engineering and Information Technology, UTS. He moved to Carnegie Mellon University, where he was a Postdoctoral Researcher with the School of Computer Science. In 2013, he came to Australia and joining the University of Queensland as a Discovery Early Career Researcher Award (DECRA) Researcher and joined UTS in 2015. His research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong. He is currently a Research Assistant Professor with The Hong Kong Polytechnic University, Hong Kong. His current research interests include image/video stylization, artistic rendering and synthesis, and creative media.

Fei Wu (Senior Member, IEEE) received the B.Sc. degree in computer science from Lanzhou University in 1996, the M.Sc. degree in computer science from the University of Macau in 1999, and the Ph.D. degree in computer science from Zhejiang University in 2002. From October 2009 to August 2010, he was a Visiting Scholar at Prof. Bin Yu's Group, University of California at Berkeley. He is currently a Qiushi Distinguished Professor with the College of Computer Science, Zhejiang University, where he is also the Vice Dean of the College of Computer Science and the Director of the Institute of Artificial Intelligence. His research interests mainly include artificial intelligence, multimedia analysis and retrieval, and machine learning. He has won various honors, such as the Award of National Science Fund for Distinguished Young Scholars of China in 2016. He is currently an Associate Editor of *Multimedia Systems* and an Editorial Member of *Frontiers of Information Technology & Electronic Engineering*.

Yueting Zhuang (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Zhejiang University, China, in 1986, 1989, and 1998, respectively. From February 1997 to August 1998, he was a Visiting Scholar with the University of Illinois at Urbana-Champaign. From 2009 to 2017, he served as the Dean of the College of Computer Science, Zhejiang University, and the Director of the Institute of Artificial Intelligence from 2006 to 2015. He is currently a Full Professor with the College of Computer Science and the Director of the MOE-Digital Library Engineering Research Center, Zhejiang University. His research interests include multimedia retrieval, artificial intelligence, cross-media computing, and digital library. He has won various awards and honors, such as the National Science Fund for Distinguished Young Scholars of China from the National Natural Science Foundation (2005), the "Chang Jiang Scholars Program" Professor of the Ministry of Education of China (2008), and the Chief Scientist of 973 Project (2012CB316400).