Synergistic Saliency and Depth Prediction for RGB-D Saliency Detection

Yue Wang[1], Yuke Li[2], James H. Elder[3], Runmin Wu[4], Huchuan Lu[1⋆], and Lu Zhang[1]

[1] Dalian University of Technology
[2] UC Berkeley
[3] York University
[4] The University of Hong Kong

Abstract. Depth information available from an RGB-D camera can be useful in segmenting salient objects when figure/ground cues from RGB channels are weak. This has motivated the development of several RGB-D saliency datasets and algorithms that use all four channels of the RGB-D data for both training and inference. Unfortunately, existing RGB-D saliency datasets are small, which may lead to overfitting and limited generalization for diverse scenarios. Here we propose a semi-supervised system for RGB-D saliency detection that can be trained on smaller RGB-D saliency datasets *without* saliency ground truth, while also make effective joint use of a large RGB saliency dataset with saliency ground truth together. To generalize our method on RGB-D saliency datasets, a novel prediction-guided cross-refinement module which jointly estimates both saliency and depth by mutual refinement between two respective tasks, and an adversarial learning approach are employed. Critically, our system does not require saliency ground-truth for the RGB-D datasets, which saves the massive human labor for hand labeling, and does not require the depth data for inference, allowing the method to be used for the much broader range of applications where only RGB data are available. Evaluation on seven RGB-D datasets demonstrates that even without saliency ground truth for RGB-D datasets and using only the RGB data of RGB-D datasets at inference, our semi-supervised system performs favorable against state-of-the-art fully-supervised RGB-D saliency detection methods that use saliency ground truth for RGB-D datasets at training and depth data at inference on two largest testing datasets. Our approach also achieves comparable results on other popular RGB-D saliency benchmarks.

Keywords: RGB-D Saliency Detection; Semi-supervised Learning; Cross Refinement; Adversarial Learning

1 Introduction

Salient Object Detection (SOD) aims to accurately segment the main objects in an image at the pixel level. It is an early vision task important for downstream

⋆ Corresponding author. Email Address: lhchuan@dlut.edu.cn

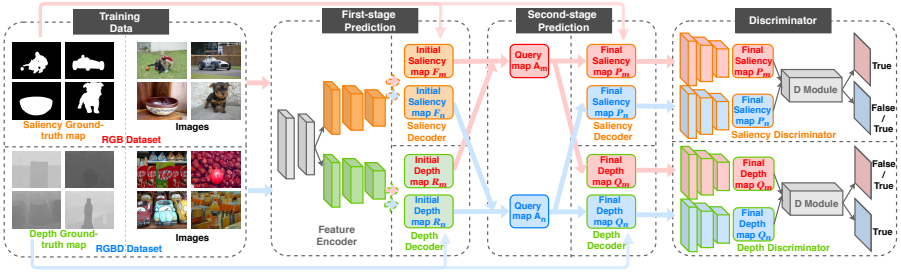

Fig. 1. Overview of our proposed method, including the first-stage prediction module, the second-stage prediction module, and our discriminator module. The input of our structure is RGB data only, and during testing, we only take RGB data from RGB-D saliency datasets to produce the saliency prediction with the Final Saliency map P.

tasks such as visual tracking [1], object detection [2], and image-retrieval [3]. Recently, deep learning algorithms trained on large ($> 10K$ image) RGB datasets like DUTS [4] have substantially advanced the state of the art. However, the problem remains challenging when figure/ground contrast is low or backgrounds are complex.

It has been observed that in these cases depth information available from an RGB-D camera can be useful in segmenting the salient objects, which are typically in front of the background [5–12]. This has motivated the development of several small RGB-D saliency datasets [5,13–18] with pixel-level hand-labeled saliency ground-truth maps for training. In order to emphasize the value of depth information, these datasets were constructed so that segmentation based only on RGB channels is difficult due to similarities in colour, texture and 2D configural cues in figure and ground (Fig. 1). Note that algorithms trained on these datasets use all four channels of the RGB-D data for both training and inference.

Unfortunately, RGB-D images are much rarer than RGB images, and existing RGB-D saliency datasets are much smaller than existing RGB saliency datasets (several hundred vs ten thousand images), which might lead to overfitting and limited generalization for diverse scenarios. In theory, one could construct a much larger RGB-D dataset with hand-labeled saliency ground truth, but this would entail specialized equipment and an enormous amount of human labor. Moreover, the existing fully-supervised methods require the additional depth map as input during inference, which limits their applications and makes them be not suitable when only RGB data are available.

This raises the question: Is it possible to make joint use of large RGB saliency datasets with hand-labeled saliency ground truth, together with smaller RGB-D saliency datasets *without* saliency ground truth, for the problem of saliency detection on RGB-D datasets? This would allow us to recruit the massive hand-labeled RGB saliency datasets that already exist while facilitating the expansion of RGB-D training datasets, since hand-labeled saliency maps for these images is not required. Perhaps an even more interesting and ambitious question is: Can

we train a semi-supervised system using these two disparate data sources such that it can perform accurate inference on the kinds of images found in RGB-D saliency datasets, *even when given only the RGB channels*? This would allow the system to be used in the much broader range of applications for which only RGB data are available.

However, the images found in RGB-D saliency datasets are statistically different from the images found in typical RGB saliency datasets since they contain more complicated background, how to make our semi-supervised model trained with RGB dataset and its saliency ground truth to be generalized well on RGB-D datasets without the saliency ground truth becomes the key challenge. To address this challenge, we propose our novel prediction-guided cross-refinement network with adversarial learning.

The system consists of three stages (Fig. 1). **Stage 1.** We build an initial prediction module with two branches: a saliency branch that takes RGB images from an RGB saliency dataset as input and is supervised with saliency ground truth; and a depth branch that takes RGB images from an RGB-D saliency dataset as input and is supervised with depth ground truth (i.e., the Depth data of the RGB-D images). **Stage 2.** Since in stage one, for each source dataset, only one branch is supervised, the statistical difference between two sources makes our initial model not generalize well on the unsupervised source, we propose our prediction-guided cross-refinement module as a bridge between two branches. The supervised branch contributes to the unsupervised one with extra information which promotes the generalization of our model. **Stage 3.** To further solve the distribution difference for two sources in this semi-supervised situation, we employ a discriminator module trained adversarially, which serves to increase the similarity in representations across sources.

Not only do we train our RGB-D saliency prediction model without the saliency ground truth from RGB-D datasets, we also do not need depth data at inference: the depth data of RGB-D images is used only as a supervisory signal during training. This makes our system usable not just for RGB-D data but for the wider range of applications where only RGB data are available. We evaluate our approach on seven RGB-D datasets and show that our semi-supervised method achieves comparable performance to the state-of-the-art fully-supervised methods, which, unlike our approach, use hand-labeled saliency ground truth for RGB-D datasets at training and use the depth data at inference.

In summary, we make two main contributions:

- We introduce a novel semi-supervised method with prediction-guided cross-refinement module and adversarial learning that effectively exploits large existing hand-labeled RGB saliency dataset, together with *unlabelled RGB-D data* to accurately predict saliency maps for RGB data from RGB-D saliency datasets. To the best of our knowledge, our paper is the first exploration of the semi-supervised method for RGB-D saliency detection.
- We show that, our semi-supervised method which does not use saliency ground truth for RGB-D datasets during training and uses only the RGB data at inference, performs favorable against existing fully-supervised meth-

ods that use saliency ground truth for RGB-D data at training and use the RGB data as well as depth data at inference on two largest RGB-D testing datasets (SIP and STEREO), and achieves comparable results on other popular RGB-D saliency benchmarks.

2 Related Work

Considering that it is still a challenge for the existing RGB saliency detections trained on RGB datasets tend to process images with complex scenarios, new RGB-D datasets with complex-scenario images and depth data are constructed to focus on this circumstance [5, 13–17]. The spatial structure information provided by depth data can be of great help for saliency detection, especially for situations like lower contrast between foreground and background. Several methods focus on RGB-D saliency detection have been proposed to achieve better performance on images with complex scenarios.

In the early stage, approaches like [11,12,14,17,19,20] use traditional methods of hand-crafted feature representations, contextual contrast and spatial prior to extract information and predict saliency maps from both RGB data and depth data in an unsupervised way. [14] proposes the first large scale RGB-D benchmark dataset and a detection algorithm which combines depth information and appearance cues in a coupled manner. More recently, supervised CNN models that extract high-level content have been found beneficial to saliency detection for complex images. Methods based on CNN structures achieve better performance on RGB-D saliency detection [5–10, 21, 22]. [21] employs two CNN networks to deal with RGB and depth data separately, and fuses the two networks on prediction level to predict the final saliency map, while [22] fuses the two networks on feature level to predict the final saliency map. [5] applies the multi-scale recurrent attention network to combine features from RGB and depth data at multiple scales, which considers both global and local information.

However, the above methods suffer from two problems. First, RGB-D datasets are rarer and the number of images in the existing RGB-D datasets is much smaller, which makes the above methods may tend to be overfitting and perform limitedly for diverse situations. Build larger RGB-D datasets for training would require not only massive labor work on labeling the pixel-level ground-truth saliency maps, but also special equipment to collect depth data. Second, the above methods demand depth data in both training and inference processes, which limits the application of RGB-D saliency detection to images with both RGB data and depth data. In this paper, we propose our semi-supervised method with prediction-guided cross-refinement module and adversarial learning to predict saliency maps for RGB-D datasets with the help of the existing RGB dataset and its saliency ground truth as well as our designed structure, we are able to train the saliency prediction model for RGB-D datasets *without accessing to their saliency ground truth*. Besides, by using depth data as an auxiliary task instead of input, it allows us to evaluate our model with only RGB data.

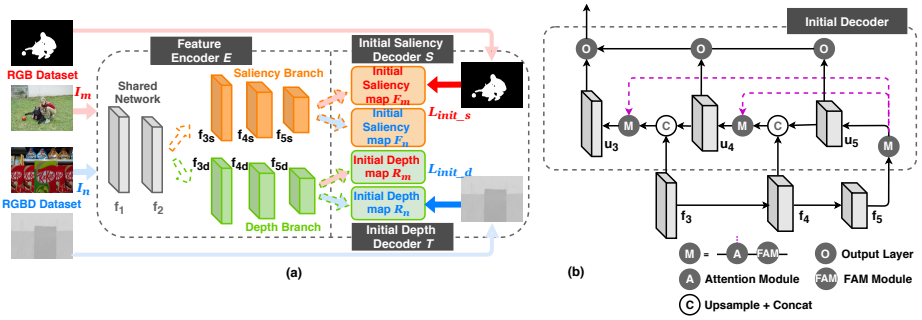

Fig. 2. (a) Illustration of our proposed first-stage initial prediction module. It outputs the initial saliency and depth prediction maps separately for both RGB and RGB-D datasets. (b) The detailed structure of our decoder, the saliency decoder and depth decoder apply the same structure.

3 Method

In this paper, we propose a novel semi-supervised approach for RGB-D saliency detection by exploiting small RGB-D saliency datasets without saliency ground truth, and large RGB saliency dataset with saliency ground truth. It contains three stages: a first-stage initial prediction module with two branches for saliency and depth tasks where each task is supervised with only one source dataset; a second-stage prediction-guided cross-refinement module which provides a bridge between the two branches for each source; and a third-stage discriminator module which further aligns representations from two sources. The overview of our proposed structure is shown in Fig. 1.

3.1 Prediction Module: The First Stage

The basic structure of our first stage prediction module showed in Fig. 2 consists of a feature encoder **E**, an initial saliency decoder **S**, and an initial depth decoder **T**. Our **E** is based on a VGG19 [23] backbone, which extracts features at five levels. Both of our decoders **S** and **T** apply the same architecture using the last three levels of features from the encoder **E**, similar to FCN [24]. To simultaneously perform two different tasks, our **E** is designed to have a two-branch structure for both saliency and depth feature representations. It has a common two-level root to constrain the model size by sharing weights on the first two levels, and it is followed by three separate layers for each branch encoding features that are passed to the respective decoders. In summary, our feature encoder **E** extracts 8 layers of features for each image: two features layers common to both saliency and depth $\{f_1, f_2\}$, three saliency-specific feature layers $\{f_{3s}, f_{4s}, f_{5s}\}$, and three depth-specific feature layers $\{f_{3d}, f_{4d}, f_{5d}\}$.

To further improve the prediction accuracy, we incorporate an extra attention module for features on each level for two decoders. We first introduce a very

basic self-attention module from the non-local block [25], which is an implementation of the self-attention form in [26]. Given a query and a key-value pair, the attention function can be described as to learn a weighted sum of values with the compatibility function of the query and key. For self-attention module, query, key, and value are set to be the same, and the weighted sum output is:

$$u = W_z(\text{softmax}(f^T W_\theta^T W_\phi f)g(f)) + f \tag{1}$$

where f is the input feature, u is the weighted sum output. W_θ, W_ϕ, $g(\cdot)$ and W_z are the function for query, key, value and weight (See [25, 26] for details).

For the highest-level feature f_5, we apply the idea of the above self-attention module (Eq. 1) which uses the f_5 itself as the query to obtain the output feature u_5. While for the feature from the other level f_L, $L \in \{4, 3\}$, it first need to combine with a higher-level output u_{L+1} by the following common practice:

$$\tilde{f}_L = conv(\text{cat}(\text{UP}(u_{L+1}), f_L)) \tag{2}$$

where $L \in \{3, 4\}$ indicates the level of feature, $\text{cat}(\cdot)$ is the concat function, $\text{UP}(\cdot)$ is the function for upsampling.

We also apply the attention module for the lower-level features. However, considering the fact that features on different levels are complementary to each other since they extract information in different resolutions, high-level features focus on global semantic information, and low-level features provide spatial details which may contain noises, we would like to select which fine details to pay attention to in low-level features with the global context. Therefore, the attention module we apply to lower-level features $\tilde{f}_L, L \in \{3, 4\}$ are based on the highest-level feature u_5 to extract meaningful details for prediction. Based on the idea of Eq. 1, we replace the query with feature u_5 and form our feature-guided attention module. The overall feature-guided attention module is as follow:

$$u_L = \begin{cases} W_{z_L}(\text{softmax}(f_L^T W_{\theta_L}^T W_{\phi_L} f_L)g_L(f_L)) + f_L & L = 5 \\ W_{z_L}(\text{softmax}(u_5^T W_{\theta_L}^T W_{\phi_L} \tilde{f}_L)g_L(\tilde{f}_L)) + \tilde{f}_L & L \in \{4, 3\} \end{cases} \tag{3}$$

where $\tilde{f}_L$ is the combined feature, and $u5$ is the updated feature on f_5.

Meanwhile, we also apply the FAM module [27] for all-level features. It is capable of reducing the aliasing effect of upsampling as well as enlarging the receptive field to improve the performance. We then apply three prediction layers on multi-level features $\{u_5, u_4, u_3\}$ and add the outputs together to form the initial prediction regarding the branch they belong to. The detailed architecture of our decoder is illustrated in Fig. 2(b).

Given an image I_m from RGB dataset with its saliency ground truth Y_m, and an image I_n from RGB-D dataset with its depth data Z_n, we can obtain their corresponding initial saliency and depth features $\{u_{3s}, u_{4s}, u_{5s}, u_{3d}, u_{4d}, u_{5d}\}_m$ and $\{u_{3s}, u_{4s}, u_{5s}, u_{3d}, u_{4d}, u_{5d}\}_n$ with the same encoder $\mathbf{E}$ and separate decoders $\mathbf{S}$, $\mathbf{T}$. The three levels of saliency features belong to image I_m will then be used to output its initial saliency maps F_m, while the three levels of depth features belong to image I_n will then be used to output its initial depth maps R_n. Since

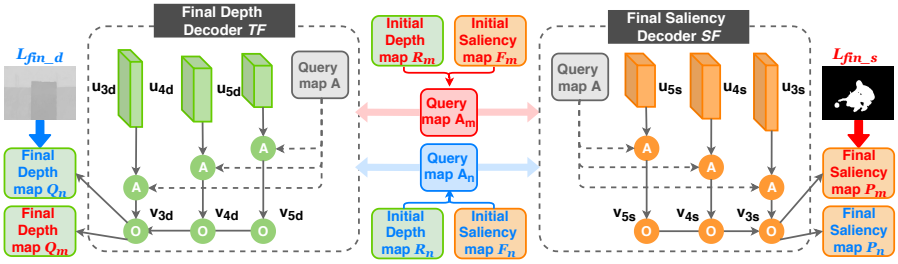

Fig. 3. Illustration of our proposed second-stage prediction module. It uses the initial saliency and depth prediction maps as the query to cross refine feature representations on both branches for RGB dataset and RGB-D dataset.

Y_m of I_m and Z_n of I_n are available, we can use them to calculate the losses of two initial maps to train our first stage prediction model:

$$\mathcal{L}_{init\_s}(\mathbf{E}, \mathbf{S}) = \mathcal{L}_{bce}(F_m, Y_m) \tag{4}$$

$$\mathcal{L}_{init\_d}(\mathbf{E}, \mathbf{T}) = \mathcal{L}_1(R_n, Z_n) \tag{5}$$

For the saliency branch, we calculate it using the binary cross-entropy loss $\mathcal{L}_{bce}$, and for the depth branch, we calculate it using the L1 loss $\mathcal{L}_1$. The overall architecture of our first-stage prediction is illustrated in Fig. 2.

3.2 Prediction Module: The Second Stage

In the first stage of our prediction module, the saliency and depth branch can only affect each other on the two shared layers in **E**. Since **S** is only supervised with images from the RGB dataset and **T** is only supervised by images from the RGB-D dataset, these two decoders may not be generalized well on the unsupervised source datasets since the difference between RGB and RGB-D saliency datasets on distribution. However, we notice that the initial depth map R_n from RGB-D dataset which provides spatial structural information can be helpful for its saliency prediction, while the initial saliency map F_m from RGB dataset can be assisted to its depth prediction since it shows the location of the important objects which draw people's attention. Therefore, to enhance the generalization for our initial module on different source datasets, we come up with an idea of using a prediction-guided cross-refinement module as a bridge to transfer information between saliency and depth branch.

Here, we build our final saliency decoder **SF** and final depth decoder **TF**, which use our designed prediction-guided method to cross refine the feature representations and initial maps from the first stage. Our prediction-guided cross-refinement method is based on the same idea of feature-guided attention module in Sec. 3.1. The detailed structure of our second stage prediction module is shown in Fig. 3.

In this stage, given features from two branches, $\{u_{3s}, u_{4s}, u_{5s}, u_{3d}, u_{4d}, u_{5d}\}$, the initial saliency map F as well as initial depth map R are used as the query

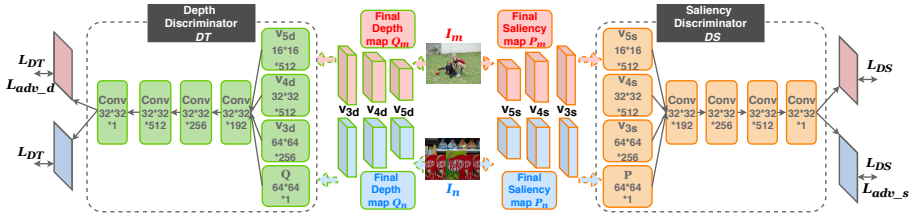

Fig. 4. Illustration of our discriminator module for adversarial learning. It has two parts, the discriminator DS deals with representations from the saliency branch, and the discriminator DT deals with representations from the depth branch.

in the attention module. We first concat F and R to form the query A since the initial F will also support the saliency branch itself to focus on the more informative spatial positions and channels in saliency representations and it is the same for R to our depth branch. And then we design a prediction-guided attention module with the following equation to update all the multi-level features from two branches.

$$v_L = W_{zn_L}(\text{softmax}(A^T W_{\theta n_L}^T W_{\phi n_L} u_L) g_{n_L}(u_L)) + u_L \tag{6}$$

where u_L represents the feature from the first stage and v_L is the updated feature, $L \in \{5, 4, 3\}$. All six features from one image $\{v_{3s}, v_{4s}, v_{5s}, v_{3d}, v_{4d}, v_{5d}\}$ will then be applied to new prediction layers specific to their tasks. For images from RGB dataset, we sum up three-level saliency outputs to get the final saliency predictions P_m, and then calculate the loss with saliency ground-truth Y_m by:

$$\mathcal{L}_{fin\_s}(\mathbf{E}, \mathbf{SF}, \mathbf{S}, \mathbf{T}) = \mathcal{L}_{bce}(P_m, Y_m) \tag{7}$$

And for images from RGB-D dataset, we also sum up all three-level depth outputs to get the final depth predictions Q_n, and calculate the loss with depth ground-truth Z_n by:

$$\mathcal{L}_{fin\_d}(\mathbf{E}, \mathbf{TF}, \mathbf{S}, \mathbf{T}) = \mathcal{L}_1(Q_n, Z_n) \tag{8}$$

3.3 Discriminator

Take the saliency branch as an example, in the first stage, the saliency prediction is only supervised by RGB saliency dataset. Even though we apply our feature-guided attention and FAM module, it can be only helpful for improving the performance for saliency detection on RGB saliency dataset itself. Due to the different distribution between RGB and RGB-D saliency datasets since images from RGB-D dataset contain more complicated background with similarities to the foreground in color and texture, the improvement on RGB dataset may also cause more noisy background to be detected as the salient region on RGB-D dataset by mistake (Fig. 5 from B to B+M). With extra depth information from RGB-D saliency dataset and the cross refinement module in our second

stage, the information from the supervised depth branch is able to transfer to the unsupervised saliency branch for RGB-D dataset, the generalization ability of our saliency prediction model on RGB-D dataset can be enhanced, such as the first row in Fig. 5 from B+M to B+M+A. While for some images which are significantly different from RGB dataset on appearance and situations as the one in second row, the effectiveness of our prediction-guided cross-refinement module may be affected. To further generalize our saliency model on RGB-D dataset, we take advantage of the adversarial learning to narrow down the distance between the representations from RGB and RGB-D dataset by adding a discriminator module (Fig. 5 from B+M+A to Ours). It could also be equally applied to the depth branch, and the detail of our discriminator module is shown in Fig. 4.

The original idea of adversarial learning is used for Generative Adversarial Network (GAN) [28], which is to generate fake images from noise to look real. It is further used in domain adaptation for image classification [29, 30], object detection [31,32] and semantic segmentation [33–36], where they train the model on source domain with easily obtained ground truth and generalize it to target domain without ground truth. The purpose of the domain adaptation is to solve the problem of domain shift due to image difference on appearance, textures, or style for two domains. The adversarial learning method uses the generator and discriminator modules to compete against each other to minimize the distance between distributions of representations on two domains, which is also suitable for our semi-supervised method to further improve its generalization ability.

Our discriminator module has two parts that respond to two task branches, discriminator **DS** is for the saliency branch, and discriminator **DT** is for the depth branch. These two discriminators are trained to distinguish representations from RGB and RGB-D dataset, and our two-stage prediction module is treated as the generator to fool the discriminators. The adversarial learning on generator and discriminators helps our prediction model to extract useful representations for saliency and depth tasks which can be generalized on both source datasets. Here, we align both latent feature representations and output prediction representations from the two datasets. For **DS**, since image I_m from RGB dataset have the saliency ground truth Y_m, we train **DS** so that the saliency feature representations $\{v_{3s}, v_{4s}, v_{5s}\}_m$ and output representation P_m can be classified as source domain label 0, while the representations $\{v_{3s}, v_{4s}, v_{5s}\}_n$ and P_n from image I_n in RGB-D dataset can be classified as target domain label 1. And we calculate the loss of **DS** by:

$$\begin{aligned}\mathcal{L}_{DS}(\mathbf{DS}) = \mathcal{L}_{bce}(\mathbf{DS}(v_{3s_m}, v_{4s_m}, v_{5s_m}, P_m), 0) \\ +\mathcal{L}_{bce}(\mathbf{DS}(v_{3s_n}, v_{4s_n}, v_{5s_n}, P_n), 1)\end{aligned} \tag{9}$$

where $\mathcal{L}_{bce}$ is the binary cross-entropy domain classification loss since the output channel of our discriminator is 1. Meanwhile, instead of predicting one value for the whole image, we obtain a patch-level output corresponding to the patch-level representations, which allows the discriminator to predict different labels for each patch, in order to encourage the system to learn the diversity of factors that determine the generalization for each spatial position.

For **DT**, depth representations $\{v_{3d}, v_{4d}, v_{5d}\}_n$ and Q_n from image I_n are supervised by depth ground-truth data Z_n, so we train **DT** to classify its representations as source domain label 0, and classify representations $\{v_{3d}, v_{4d}, v_{5d}\}_m$ and Q_m from I_m as target domain label 1. The loss for **DT** is calculated by:

$$\begin{aligned}\mathcal{L}_{DT}(\mathbf{DT}) = \mathcal{L}_{bce}(\mathbf{DT}(v_{3d_n}, v_{4d_n}, v_{5d_n}, Q_n), 0) \\ +\mathcal{L}_{bce}(\mathbf{DT}(v_{3d_m}, v_{4d_m}, v_{5d_m}, Q_m), 1)\end{aligned} \tag{10}$$

To fool **DS**, our prediction model is trained to learn saliency representations $\{v_{3s}, v_{4s}, v_{5s}\}_n, P_n$ from I_n which can be classified as source domain in **DS**. The adversarial loss for saliency branch can be calculated as:

$$\mathcal{L}_{adv\_s}(\mathbf{E}, \mathbf{SF}, \mathbf{S}, \mathbf{T}) = \mathcal{L}_{bce}(\mathbf{DS}(v_{3s_n}, v_{4s_n}, v_{5s_n}, P_n), 0) \tag{11}$$

For **DT**, our prediction model is trained to learn depth representations $\{v_{3d}, v_{4d}, v_{5d}\}_m, Q_m$ from I_m which can be classified as source domain:

$$\mathcal{L}_{adv\_d}(\mathbf{E}, \mathbf{TF}, \mathbf{S}, \mathbf{T}) = \mathcal{L}_{bce}(\mathbf{DT}(v_{3d_m}, v_{4d_m}, v_{5d_m}, Q_m), 0) \tag{12}$$

3.4 Complete Training Loss

To summarize, the complete training process includes losses for our prediction model, which combines the initial saliency prediction loss for I_m (Eq. (4)), the initial depth prediction loss for I_n (Eq. (5)), the final saliency prediction loss for I_m (Eq. (7)), the final depth prediction loss for I_n (Eq. (8)), the adversarial loss of saliency branch for I_n (Eq. (11)), the adversarial loss of depth branch for I_m (Eq. (12)); and the losses for saliency and depth discriminators (Eq. (9), (10)),

$$\min_{\mathbf{DS},\mathbf{DT}} \mathcal{L}_{DS} + \mathcal{L}_{DT} \tag{13}$$

$$\begin{aligned}\min_{\mathbf{E},\mathbf{SF},\mathbf{TF},\mathbf{S},\mathbf{T}} \quad & \lambda_s \mathcal{L}_{fin\_s} + \lambda_d \mathcal{L}_{fin\_d} \\ + \quad & \lambda_{init}\lambda_s \mathcal{L}_{init\_s} + \lambda_{init}\lambda_d \mathcal{L}_{init\_d} \\ + \quad & \lambda_{adv\_s} \mathcal{L}_{adv\_s} + \lambda_{adv\_d} \mathcal{L}_{adv\_d}\end{aligned} \tag{14}$$

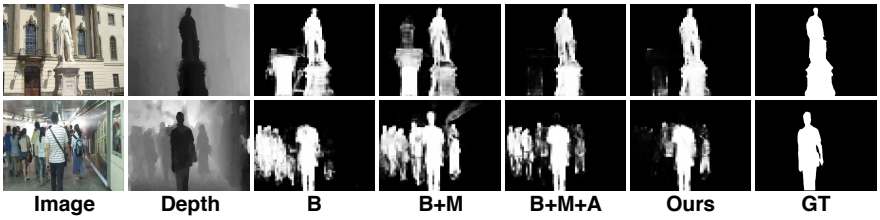

Fig. 5. Visual examples for ablation study. See Section 4.4 for the definition of each subset model.

4 Experiments

In this section, we evaluate our method and present the experimental results. First, we introduce the benchmark datasets and some implementation details of our network architecture. Then, we discuss the effectiveness of our method by comparison with the state-of-art methods and the ablation study.

4.1 Datasets and Evaluation Metrics

We evaluate our proposed method on seven widely used RGB-D saliency datasets including NJUD [13], NLPR [14], LFSD [15], STEREO [16], RGBD135 [17], SIP [18], and DUT-D [5]. To train our model, we use the DUTS [4], an RGB saliency dataset contains 10553 images with saliency ground truth for training the saliency branch. And for our depth branch, for a fair comparison, we use the selected 1485 NJUD images and 700 NLPR images as in [5] without saliency ground truth as our RGB-D training set. We then evaluate our model on 797 images in STEREO, 929 images in SIP (These two datasets contain the largest number of images for the test split); and other testing datasets including 100 images in LFSD, 135 images in RGBD135, 400 images in DUT-D testing set; as well as the remaining 500 testing images in NJUD, 300 testing images in NLPR.

For quantitative evaluation, we adopt four widely used evaluation metrics including F-measure(F_m) [37], mean absolute error (MAE) [38], S-measure(S_m) [39] and E-measure(E_m) [40]. In this paper, we report the average F-measure value as F_m which is calculated by the mean of the precision and recall. For MAE, the lower value indicates the method is better, while for all other metrics, the higher value indicates the method is better.

4.2 Implementation Details

We apply PyTorch for our implementation using two GeForce RTX 1080 Ti GPU with 22 GB memory. For our prediction model, we use VGG19 [23] pre-trained model as the backbone. And for the discriminator, we first apply one convolution layer for each input feature/prediction and concat the latent representations, then apply four convolution layers to output the one-channel classification result. We apply ADAM [41] optimizer for both two-stage prediction module and discriminator module, with the initial learning rate setting to 1e-4 and 5e-5. We set $\lambda_s = 1.75, \lambda_d = 1.0, \lambda_{init} = 0.2, \lambda_{adv\_s} = 0.002, \lambda_{adv\_d} = 0.001$ to focus more on the saliency branch and the second stage. All the input images are resized to 256×256 pixels.

4.3 Comparison with state-of-the-art methods

We compare our method with 9 state-of-the-art RGB-D saliency detection methods including 7 RGB-D deep learning methods: DMRA [5], CPFP [6], TANet [7], MMCI [8], PCANet [9], CTMF [21], DF [10], and 2 RGB-D traditional methods: DCMC [11], CDCP [12]. The performance of our method compared with

Table 1. Results on different datasets. We highlight the best two result in each column in red and blue.

	DUT-D				STEREO				SIP				RGBD135			
	MAE	F_m	S_m	E_m	MAE	F_m	S_m	E_m	MAE	F_m	S_m	E_m	MAE	F_m	S_m	E_m
DMRA	0.048	0.883	0.887	0.930	0.047	0.868	0.886	0.934	0.088	0.815	0.800	0.858	0.030	0.867	0.899	0.944
CPFP	0.100	0.735	0.749	0.815	0.054	0.827	0.871	0.902	0.064	0.819	0.850	0.899	0.038	0.829	0.872	0.927
TANet	0.093	0.778	0.808	0.871	0.059	0.849	0.877	0.922	0.075	0.809	0.835	0.894	0.046	0.795	0.858	0.919
MMCI	0.112	0.753	0.791	0.856	0.080	0.812	0.856	0.894	0.086	0.795	0.833	0.886	0.065	0.762	0.848	0.904
PCANet	0.100	0.760	0.801	0.863	0.061	0.845	0.880	0.918	0.071	0.825	0.842	0.900	0.050	0.774	0.843	0.912
CTMF	0.097	0.792	0.831	0.883	0.087	0.786	0.853	0.877	0.139	0.684	0.716	0.824	0.055	0.778	0.863	0.911
DF	0.145	0.747	0.729	0.842	0.142	0.761	0.763	0.844	0.185	0.673	0.653	0.794	0.131	0.573	0.685	0.806
DCMC	0.243	0.405	0.499	0.712	0.150	0.762	0.745	0.838	0.186	0.645	0.683	0.787	0.196	0.234	0.469	0.676
CDCP	0.159	0.633	0.687	0.794	0.149	0.681	0.727	0.801	0.224	0.495	0.595	0.722	0.120	0.594	0.709	0.810
Ours	0.057	0.878	0.885	0.935	0.045	0.878	0.893	0.936	0.052	0.856	0.880	0.922	0.031	0.864	0.890	0.927

the state-of-the-art methods on each evaluation metric is showed in Table 1 and Table 2. For a fair comparison, the saliency maps of the above methods we use are directly provided by authors, or predicted by their released codes. We apply the same computation of the evaluation metrics to all the saliency maps.

For all the listed latest RGB-D methods based on CNNs-based structure, they all require depth data as input for both training and inference, and they use RGB-D saliency ground-truth maps to train the model in a fully-supervised way. Therefore, they can achieve a good performance on all the datasets. For RGB-D traditional methods, they use manually designed cues to calculate the saliency prediction in an unsupervised way, and they perform worse compared with the CNN-based fully-supervised RGB-D methods. With the help of images and saliency ground-truth maps from RGB datasets, our semi-supervised method does not require access to any saliency ground-truth maps for images in RGB-D datasets during training, and we only require the RGB data without depth data at inference since we use the depth data as a supervisory signal during training.

The quantitative results show that, for the two largest testing SIP and STEREO datasets containing the largest number of images for testing, our semi-supervised method can achieve better results, which indicates that our method may generalize better on diverse scenario even without having access to saliency ground truth for RGB-D datasets. It can also demonstrate that useful information can be obtained from a larger RGB saliency dataset and generalized to RGB-D saliency datasets by our designed approach, despite that the images from these two source datasets have considerable difference on appearance since RGB-D datasets focus on images with more complicated background. For other datasets with a smaller number of images for testing such as DUT-D and RGBD135, we are also able to reach comparable results with DMRA which are better than other methods. We may perform slightly worse on two specific datasets, NJUD and NLPR, since all other fully-supervised methods use the saliency ground-truth maps from these two datasets during training. However, we still manage to be comparable with the state-of-art methods on these two datasets. To better demonstrate the advantage of our method, we also present some qualitative saliency examples in Fig. 6.

Table 2. Results on different datasets. We highlight the best two result in each column in red and blue.

	LFSD				NJUD				NLPR			
	MAE	F_m	S_m	E_m	MAE	F_m	S_m	E_m	MAE	F_m	S_m	E_m
DMRA	0.076	0.849	0.847	0.899	0.051	0.872	0.885	0.920	0.031	0.855	0.898	0.942
CPFP	0.088	0.813	0.828	0.867	0.053	0.837	0.878	0.900	0.038	0.818	0.884	0.920
TANet	0.111	0.794	0.801	0.851	0.061	0.844	0.878	0.909	0.041	0.796	0.886	0.916
MMCI	0.132	0.779	0.787	0.840	0.079	0.813	0.859	0.882	0.059	0.730	0.856	0.872
PCANet	0.112	0.794	0.800	0.856	0.059	0.844	0.877	0.909	0.044	0.795	0.874	0.916
CTMF	0.120	0.781	0.796	0.851	0.085	0.788	0.849	0.866	0.056	0.724	0.860	0.869
DF	0.142	0.810	0.786	0.841	0.151	0.744	0.735	0.818	0.100	0.683	0.769	0.840
DCMC	0.155	0.815	0.754	0.842	0.167	0.715	0.703	0.796	0.196	0.328	0.550	0.685
CDCP	0.199	0.634	0.658	0.737	0.182	0.618	0.672	0.751	0.115	0.592	0.724	0.786
Ours	0.090	0.823	0.830	0.879	0.055	0.852	0.878	0.909	0.044	0.809	0.875	0.915

4.4 Ablation Study

To demonstrate the impact of each component in our overall method, we conducted our ablation study by evaluating the following subset models:

1) B: Our baseline, a simple saliency detection model directly trained by RGB saliency dataset with only multi-level fusion in the first stage.

2) B + M: Only trained by RGB saliency dataset while adding the FAM module and our feature-guided attention module in the first stage.

3) B + M + A: Adding the depth branch trained by RGB-D saliency datasets and the second stage cross-refinement prediction with the prediction-guided attention module.

4) Ours: Our overall structure with the discriminator module.

Our ablation study is evaluated on three RGB-D datasets and the result is showed in Table 3. We also include some visual examples in Fig. 5. It indicates that our baseline model B provides a good initial prediction with the saliency branch trained by the RGB dataset. By adding feature-guided attention module which helps to focus on more informative spatial positions and channels, and the FAM module which enlarges the receptive field, B+M further improves performance by helping saliency detection on RGB saliency dataset. However, the trained B+M module may not be generalized well on RGB-D saliency detection due to the different distribution between RGB and RGB-D saliency datasets. It may perform badly on images with more complicated background (Fig. 5).

To improve the generalization ability of our model on RGB-D datasets, we then add depth branch and the second-stage prediction-guided cross-refinement

Table 3. Ablation Study on our proposed method. We highlight the best result in each column in red.

	NJUD				NLPR				STEREO			
	MAE	F_m	S_m	E_m	MAE	F_m	S_m	E_m	MAE	F_m	S_m	E_m
B	0.064	0.809	0.862	0.876	0.052	0.774	0.858	0.891	0.053	0.835	0.877	0.908
B+M	0.060	0.818	0.876	0.887	0.050	0.791	0.868	0.903	0.053	0.854	0.889	0.921
B+M+A	0.055	0.840	0.878	0.900	0.047	0.807	0.873	0.910	0.050	0.868	0.888	0.928
Ours	0.055	0.852	0.878	0.909	0.044	0.809	0.875	0.915	0.045	0.878	0.893	0.936

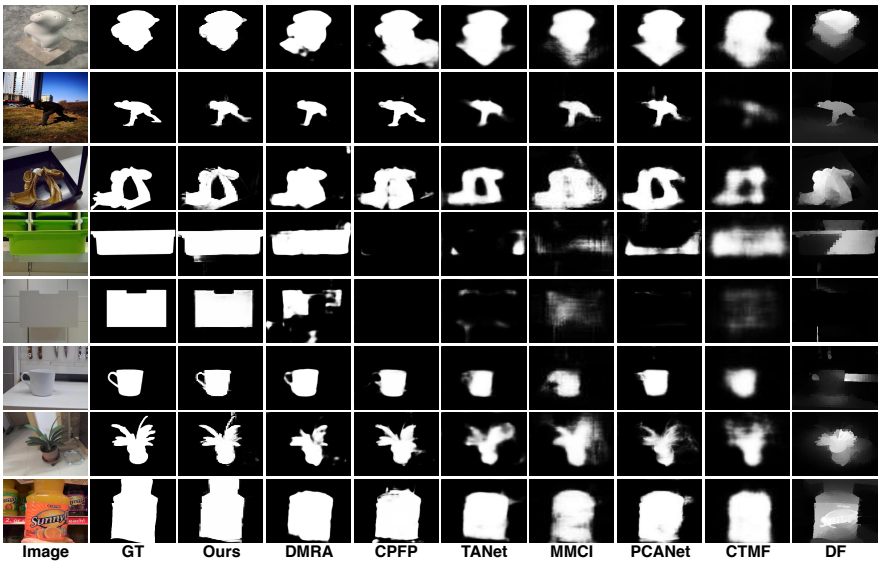

Fig. 6. Visual comparison of our method with the state-of-art methods.

module to utilize depth data with spatial structure information for RGB-D images to form B+M+A module. To further help the generalization of our model on some images from RGB-D datasets which have significant difference with images from RGB dataset, we also add the discriminator module by adversarial learning to align the representations on two source datasets for each branch to form our final model. Table 3 also proves the effectiveness of each module.

5 Conclusions

In this paper, we propose a novel semi-supervised method for RGB-D saliency detection with a synergistic saliency and depth prediction way to deal with the small number of existing RGB-D saliency datasets without constructing a new dataset. It allows us to exploit larger existing hand-labeled RGB saliency datasets, avoid using saliency ground-truth maps from RGB-D datasets during training, and require only RGB data without depth data at inference. The system consists of three stages: a first-stage initial prediction module to train two separate branches for saliency and depth tasks; a second-stage prediction-guided cross-refinement module and a discriminator stage to further improve the generalization on RGB-D dataset by allowing two branches to provide complementary information and the adversarial learning. Evaluation on seven RGB-D datasets demonstrates the effectiveness of our method, by performing favorable against the state-of-art fully-supervised RGB-D saliency methods on two largest RGB-D saliency testing datasets, and achieves comparable results on other popular RGB-D saliency detection benchmarks.

References

1. Lee, H., Kim, D.: Salient Region-based Online Object Tracking. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2018) 1170–1177
2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: International Conference on Machine Learning. (2015) 2048–2057
3. He, J., Feng, J., Liu, X., Cheng, T., Lin, T.H., Chung, H., Chang, S.F.: Mobile Product Search with Bag of Hash Bits and Boundary Reranking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2012) 3005–3012
4. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to Detect Salient Objects with Image-level Supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 136–145
5. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced Multi-Scale Recurrent Attention Network for Saliency Detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7254–7263
6. Zhao, J.X., Cao, Y., Fan, D.P., Cheng, M.M., Li, X.Y., Zhang, L.: Contrast Prior and Fluid Pyramid Integration for RGBD Salient Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3927–3936
7. Chen, H., Li, Y.: Three-stream Attention-aware Network for RGB-D Salient Object Detection. IEEE Transactions on Image Processing **28** (2019) 2825–2835
8. Chen, H., Li, Y., Su, D.: Multi-modal Fusion Network with Multi-scale Multi-path and Cross-modal Interactions for RGB-D Salient Object Detection. Pattern Recognition **86** (2019) 376–385
9. Chen, H., Li, Y.: Progressively Complementarity-aware Fusion Network for RGB-D Salient Object Detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 3051–3060
10. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: RGBD Salient Object Detection via Deep Fusion. IEEE Transactions on Image Processing **26** (2017) 2274–2285
11. Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C.: Saliency Detection for Stereoscopic Images Based on Depth Confidence Analysis and Multiple Cues Fusion. IEEE Signal Processing Letters **23** (2016) 819–823
12. Zhu, C., Li, G., Wang, W., Wang, R.: An Innovative Salient Object Detection using Center-dark Channel Prior. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2017) 1509–1515
13. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth Saliency based on Anisotropic Center-surround Difference. In: 2014 IEEE International Conference on Image Processing (ICIP), IEEE (2014) 1115–1119
14. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD Salient Object Detection: A Benchmark and Algorithms. In: European Conference on Computer Vision, Springer (2014) 92–109
15. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency Detection on Light Field. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 2806–2813
16. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging Stereopsis for Saliency Analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2012) 454–461

17. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth Enhanced Saliency Detection Method. In: Proceedings of international conference on internet multimedia computing and service. (2014) 23–27
18. Fan, D.P., Lin, Z., Zhao, J.X., Liu, Y., Zhang, Z., Hou, Q., Zhu, M., Cheng, M.M.: Rethinking RGB-D Salient Object Detection: Models, Datasets, and Large-scale Benchmarks. arXiv preprint arXiv:1907.06781 (2019)
19. Zhu, C., Li, G., Guo, X., Wang, W., Wang, R.: A Multilayer Backpropagation Saliency Detection Algorithm based on Depth Mining. In: International Conference on Computer Analysis of Images and Patterns, Springer (2017) 14–23
20. Zhu, C., Li, G.: A Three-pathway Psychobiological Framework of Salient Object Detection using Stereoscopic Technology. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2017) 3008–3014
21. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: CNNs-based RGB-D Saliency Detection via Cross-view Transfer and Multiview Fusion. IEEE transactions on cybernetics **48** (2017) 3171–3183
22. Zhu, C., Cai, X., Huang, K., Li, T.H., Li, G.: PDNet: Prior-model Guided Depth-enhanced Network for Salient Object Detection. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE (2019) 199–204
23. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-scale Image Recognition. In: ICLR. (2015)
24. Shelhamer, E., Long, J., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39** (2017) 640–651
25. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7794–7803
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is All You Need. In: Advances in Neural Information Processing Systems. (2017) 5998–6008
27. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A Simple Pooling-based Design for Real-time Salient Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3917–3926
28. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems. (2014) 2672–2680
29. Hu, L., Kan, M., Shan, S., Chen, X.: Duplex Generative Adversarial Network for Unsupervised Domain Adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1498–1507
30. Zhang, Y., Tang, H., Jia, K., Tan, M.: Domain-symmetric Networks for Adversarial Domain Adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5031–5040
31. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain Adaptive Faster R-CNN for Object Detection in the Wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 3339–3348
32. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak Distribution Alignment for Adaptive Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 6956–6965
33. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to Adapt Structured Output Space for Semantic Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7472–7481

34. Luo, Y., Liu, P., Guan, T., Yu, J., Yang, Y.: Significance-aware Information Bottleneck for Domain Adaptive Semantic Segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 6778–6787
35. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a Closer Look at Domain Shift: Category-level Adversaries for Semantics Consistent Domain Adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2507–2516
36. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2517–2526
37. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned Salient Region Detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2009) 1597–1604
38. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient Object Detection: A Benchmark. IEEE Transactions on Image Processing **24** (2015) 5706–5722
39. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A New Way to Evaluate Foreground Maps. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 4548–4557
40. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment Measure for Binary Foreground Map Evaluation. arXiv preprint arXiv:1805.10421 (2018)
41. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: ICLR. (2015)