Synergistic saliency and depth prediction for RGB-D saliency detection

Yue Wang¹, Yuke Li², James H Elder², Huchuan Lu¹, Runmin Wu³, and

¹ Dalian University of Technology
² York University
³ 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, leading to overfitting and poor generalization. Here we demonstrate a system for RGB-D saliency detection that makes effective joint use of large RGB saliency datasets with hand-labelled saliency ground truth together, and smaller RGB-D saliency datasets without saliency ground truth. This novel prediction-guided cross-refinement network is trained to jointly estimate both saliency and depth, allowing mutual refinement between feature representations tuned for the two respective tasks. An adversarial stage resolves domain shift between RGB and RGB-D saliency datasets, allowing representations for saliency and depth estimation to be aligned on either. Critically, our system does not require saliency ground-truth for the RGB-D datasets, making it easier to expand these datasets for training, and does not require the D channel for inference, allowing the method to be used for the much broader range of applications where only RGB data are available. Evaluation on seven RGBD datasets demonstrates that, without using hand-labelled saliency ground truth for RGB-D datasets and using only the RGB channels of these datasets at inference, our system achieves performance that is comparable to state-of-the-art methods that use hand-labelled saliency maps for RGB-D data at training and use the depth channels of these datasets at inference.

Keywords: RGB-D Saliency Detection, Prediction-Guided 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 tasks such as visual tracking [19], object detection [36], and image-retrieval [14]. Recently, deep learning algorithms trained on large (>10K image) RGB datasets like DUTS [34] have substantially advanced the state of the art. However, they
Fig. 1. Overview of our proposed method, including the first-stage prediction module, the second-stage prediction module, and our discriminator module.

Still remain 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 [4–6,9,27,28,38,42]. This has motivated the development of several small RGB-D saliency datasets [8,11,16,20,25–27] 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 leads to overfitting and poor generalization. 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 labour.

This raises the question: Is it possible to make joint use of large RGB saliency datasets with hand-labelled 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-labelled RGB saliency datasets that already exist, while facilitating the expansion of RGB-D training datasets, since hand-labelled saliency maps for these images is not required. Perhaps an even more interesting and ambitious question is: Can we train a 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.

As mentioned, one of the key challenges in this objective is domain shift: the images found in RGB-D saliency datasets are statistically different from the images found in typical RGB saliency datasets, and thus the cues that are most effective for saliency detection differ for the two domains.
To address this challenge, we propose a novel prediction-guided cross-refinement network trained to jointly estimate both saliency and depth, and which allows mutual refinement between feature representations tuned for the two respective tasks.

The system consists of three stages (Fig. 1). The first stage is a prediction model with two branches: 1) a saliency branch that takes RGB images from an RGB saliency dataset as input and is supervised with ground-truth saliency maps, and 2) a depth branch that takes RGB images from an RGB-D saliency dataset as input and is supervised with ground-truth depth maps (i.e., the D channel of the RGB-D images). This first stage provides us initial predictions for both tasks.

In our second stage, these initial predictions are used as queries to guide cross-refinement on both branches. In particular, we use attentional modulation across modalities to refine the initial feature representations by focusing on the more informative spatial positions and channels. Here the initial predictions from the saliency branch provide segmentation information based on RGB image features that can inform depth estimation, while the initial predictions from the depth branch provide depth information that can inform saliency detection.

Note that while our feature representations are refined by predictions from both tasks, each task is only supervised by one dataset. As noted above, there are significant domain shifts in RGB imagery and scenes between RGB and RGB-D datasets, and clearly these differences will be reflected in the saliency and depth maps produced by our second stage, which will limits generalization. To address this problem, we employ a third discriminator stage trained adversarially to discriminate whether saliency or depth maps originate from RGB or RGB-D datasets.

Note also that since the depth channel of RGB-D images is used only as a supervisory signal during training, at inference, the system estimates both saliency maps and depth maps based only on RGB channels. 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. We show that, without using hand-labelled saliency ground truth for RGB-D datasets and using only the RGB channels of these datasets at inference, our system achieves performance that is comparable to state-of-the-art methods that use hand-labelled saliency maps for RGB-D data at training and use the depth channels of these datasets at inference.

In summary, we make three main contributions:

– We introduce a novel prediction-guided cross-refinement model with adversarial learning that effectively exploits large existing hand-labelled RGB saliency dataset together with unlabelled RGB-D data to accurately predict saliency maps for RGB images from RGB-D saliency datasets. To the best of our knowledge, our paper is the first to propose an adversarial method for RGB-D saliency detection to avoid using saliency ground-truth maps from RGB-D datasets.
– We show that, without using hand-labelled saliency ground truth for RGB-D datasets and using only the RGB channels of these datasets at inference, our system achieves performance that is comparable to state-of-the-art methods that use hand-labelled saliency maps for RGB-D data at training and use the depth channels of these datasets at inference.
– Since we do not rely upon the depth channel at inference, the system can also be used for the much broader range of applications where only RGB data are available.

2 Related Work

2.1 RGB-D Saliency Object Detection

Considering that the existing RGB saliency detections trained on RGB datasets tend to fail on images with complex scenarios, Considering that it is still a challenge for the existing RGB saliency detections trained on RGB datasets to process images with complex scenarios, new RGB-D datasets with complex-scenario images and depth data are built to focus on this circumstance [8,16,20,25,27,40]. The spatial structure information provided by depth data can be of great help for saliency detection especially for the situations like lower contrast between foreground and background. Several methods focus on RGB-D saliency detection that has been proposed to achieve better performance on images with complex scenarios.

In the early stage, approaches like [8,9,26,40,42] use traditional methods of hand-crafted feature representations, contextual contrast and spatial prior to extract information and predict saliency map from both RGB data and depth data in an unsupervised way. With the development of deep learning networks, CNNs-based model which extracts high-level content information is beneficial to the saliency detection on images with complex scenarios with fully supervised training strategy. Methods based on CNN structures achieve better performance on RGB-D saliency detection [4,6,13,27,28,38,39]. [13] builds two CNN network to predict saliency maps from RGB data and depth data and fuse the two networks on prediction level, while [39] builds two CNN networks to extract feature from RGB data and depth data and fuse the two branch on feature-level to predict saliency map. [38] enhances the depth clues with a novel contrast-enhanced net to further combine it to feature representations from RGB data. [27] apply the multi-scale recurrent attention network to combine features from RGB and depth data with multiple scales, which takes both global information and local information into consideration.

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 tend to be overfitting and perform poorly in various situations. And the work to build larger RGBD datasets for training requires not only massive labor work on labeling the pixel-level ground-truth saliency maps, but also special equipment to collect depth data. Second,
Fig. 2. Illustration of our proposed first-stage prediction module. It outputs the initial saliency and depth prediction maps separately for both RGB and RGBD datasets.

all 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 prediction guided cross refinement model with adversarial learning to predict saliency maps for RGB-D datasets in a weakly supervised way. With the help of the existing RGB dataset and our designed structure, we are able to train the saliency prediction model for RGB-D dataset without accessing to its ground-truth saliency maps. Besides, we use depth data as an auxiliary task instead of input, which allows us to evaluate our model with only RGB data.

3 Method

3.1 The Overall Architecture

In this paper, we propose a novel approach that handles both saliency and depth tasks. It contains three stages: a two-stage prediction module and a discriminator module. In the first stage of our prediction module, it outputs both initial saliency and depth prediction maps separately for both RGB and RGB-D datasets supervised with the RGB’s ground-truth saliency maps and RGBD’s depth data. In the second stage, we use the initial saliency and depth maps as queries for cross refinement on feature representations and output the final saliency and depth prediction maps. Our prediction module is followed by the discriminator module, where we use adversarial learning to solve the problem of domain shift by aligning representations from two sources. The overview of our proposed structure is shown in Figure 1.

3.2 Prediction Module: The First Stage

The basic structure of our first stage prediction module includes a feature encoder $E$, an initial saliency decoder $S$, an initial depth decoder $T$. Our feature encoder $E$ is based on a VGG19 [30] backbone, which extracts feature in five
levels, denoted as \( \{f_1, f_2, f_3, f_4, f_5\} \). For simultaneously dealing with two different tasks, our module is designed with two branches for both saliency and depth feature representations. These two branches share the weights on the first two levels and learn their specific weights on the following levels. Therefore, we can extract 8 features in our feature encoder \( E \): two common features for both saliency and depth \( \{f_1, f_2\} \), three saliency specific features \( \{f_3s, f_4s, f_5s\} \), and three depth specific features \( \{f_3d, f_4d, f_5d\} \).

With the above multi-level features, we then generate both saliency and depth predictions in our designed decoders \( S \) and \( T \). Both of our decoders are based on the U-net [29] to utilize multi-level features. To further improve the performance of predictions, we add an extra attention module for features on each level. For the highest-level feature \( f_5 \), we apply the self-attention module with a basic non-local block [35], which is an implementation of the self-attention form in [32]. 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 according to [35], the weighted sum output is:

\[
u = \text{softmax}(f^T W^T \theta W \phi f)g(f) + f \tag{1}\]

where \( f \) is the input feature, \( u \) is the weighted sum output, \( W_\theta, W_\phi, \) and \( g(\cdot) \) are the function for query, key and value.

After updating our highest-level feature \( f_5 \) with self-attention module, the obtain weighted sum output \( u_5 \) will further be combined with lower-level features by the following common practice:

\[
\hat{f}_{L-1} = \text{conv} (\text{cat}(U_\text{P}(u_5), f_{L-1})) \tag{2}
\]

where \( L \) indicates the level of feature, \( \text{cat}(\cdot) \) is the concat function, \( \text{UP}(\cdot) \) is the function for upsampling. Features on different levels are complementary to each other since they extract information in different resolutions. High-level features focus on global semantic information, while low-level features further provide spatial details. However, the detail information from lower-level features are redundant, Therefore, we then apply an attention module based on the highest-level feature \( u_5 \) to the lower-level features and extract the meaningful details for prediction. Based on the idea of the self-attention module, we replace the query with feature \( u_5 \), and form our feature-guided attention module:

\[
u = \text{softmax}(u_5^T W^T \theta W \phi \hat{f})g(\hat{f}) + \hat{f} \tag{3}\]

where \( \hat{f} \) is the combined feature on level 4 and 3, and \( u_5 \) is the updated features on level 5. We then use the same function Eq. (2) for combining the lower-level features. Meanwhile, we also apply the FAM module [21] for all-level features. It is capable of reducing the aliasing effect of upsampling as well as enlarging the receptive field, and can be helpful to improve the performance. We then apply three classifiers 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.
Given an image $I_m$ from RGB datasets with its saliency ground-truth map $Y_m$, and an image $I_n$ from RGB-D datasets 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 $E$. The three levels of saliency features belong to image $I_m$ will then be used in the saliency decoder $S$ to output its initial saliency maps $F_m$, while the three levels of depth features belong to image $I_n$ will then be used in the depth decoder $T$ to output its initial depth maps $R_n$. Since the saliency ground-truth map $Y_m$ of $I_m$ and the depth data $Z_n$ of $I_n$ are available, we can use them to calculate the losses of two initial maps to train our first stage prediction model:

$$L_{\text{init.s}}(E, S) = - \sum_{h,w} H,W \sum_{c \in \{0,1\}} Y_m^{(h,w,c)} \log(F_m^{(h,w,c)})$$

$$L_{\text{init.d}}(E, T) = \sum_{h,w} H,W |R_n^{(h,w)} - Z_n^{(h,w)}|$$

where $H, W$ are the size of images, for saliency branch, we calculate it using the binary cross-entropy loss, and for depth branch, we calculate it using the L1 loss. The detailed architecture of this first stage prediction model is illustrated in Figure 2.

### 3.3 Prediction Module: The Second Stage

In the first stage of our prediction module, the saliency branch and depth branch can only affect each other on the shared layers in the encoder $E$, since the initial saliency decoder $S$ is only supervised with images from the RGB dataset, and the initial depth decoder $T$ is only supervised by images from the RGB-D dataset only. Even though we do not have the ground-truth saliency maps for the RGB-D dataset, our initial depth maps $R_n$ which provides spatial structural information can be helpful for the saliency prediction. Meanwhile, the initial saliency maps $F_m$ on RGB images can also be assisted to the depth prediction since the initial saliency maps show the location of the important objects which draw people’s attention. Furthermore, the initial saliency maps will also support the saliency...
Fig. 4. Illustration of our discriminator module for adversarial learning. It has two parts, the discriminator DS deal with representations from saliency branch, and the discriminator DT deal with representations from depth branch.

branch itself to focus on the more informative spatial positions and channels in saliency representations and it is the same for our depth branch.

Therefore, 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. 3.2. The detailed structure of our second stage prediction module is shown in Figure 3.

In this stage, given the 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 in the attention module. We first concat F and R to form the query A, and then we design a prediction-guided attention module with the following equation to update all the multi-level features from two branches.

$$v = \text{softmax}(A^T W^T W \phi u)g(u) + u$$  \hspace{1cm} (6)

where u represents the features from the first stage and v is the updated features. All the six features from one image \{v_3s, v_4s, v_5s, v_3d, v_4d, v_5d\} will then be applied to new classifiers 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:

$$\mathcal{L}_{fin,s}(E, SF, S, T) = - \sum_{h,w} \sum_{c \in \{0,1\}} Y_m^{(h,w,c)} \log(P_m^{(h,w,c)})$$  \hspace{1cm} (7)

And for images from RGBD 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:

$$\mathcal{L}_{fin,d}(E, TF, S, T) = \sum_{h,w} |Q_n^{(h,w)} - Z_n^{(h,w)}|$$  \hspace{1cm} (8)

3.4 Discriminator

The original idea of adversarial learning method is first used for Generative Adversarial Network (GAN) [12], which is to generate fake images from noise to
look real. It is further used in the area of domain adaptation for applications like image classification [15, 37], object detection [7, 17], person re-identification [2]. In these cases, the images from source domain dataset are easier to obtain the ground-truth label, but has significant difference with images from target domain dataset on appearance, textures or image style. This kind of difference makes the model trained on source domain dataset limit its generalization on the target domain dataset. With the help of the adversarial learning method, [22, 23, 31, 33] are capable of obtaining a semantic segmentation model that performs well on target domain (real-world images) by training on source domain (synthetic datasets). The purpose of the domain adaptation is to minimize the distances between distributions of representations in prediction space or feature space on two domains, and the adversarial learning methods use the generator and discriminator modules to compete against each other for realizing the domain adaptation.

In our two-stage prediction module, the saliency branch is supervised by the saliency ground-truth maps from RGB datasets, and the depth branch is supervised by depth data from RGB-D datasets. The difference between two datasets on appearance and situations would affect the generalization of the prediction model since the saliency branch focuses on predicting saliency maps with the distributions of RGB datasets, and the depth branch focuses on predicting depth maps for the situations in RGBD datasets, which makes both decoders SF, TF difficult to generalize on images with dataset with other distribution. To solve the above problem, we take advantage of the adversarial learning method to narrow down the distance between the representations from RGB dataset and RGB-D dataset by adding an extra discriminator module. The detail implement of our discriminator module is shown in Figure 4.

Our discriminator module has two parts respond to two tasks 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 datasets, and our two-stage prediction module is treated as the generator to fool the discriminators. The adversarial learning on generator and discriminator helps our prediction model to extract useful representations for saliency and depth tasks based on high contextual semantic information, which is unified for images in both datasets instead of using texture information which can be unique for different datasets. In this paper, we align the distances on both latent feature representations and output representations from the two datasets. For DS, since image $I_m$ from RGB dataset have the saliency ground truth $Y_m$, we assign the saliency feature representations $\{v3s, v4s, v5s\}_m$ and output representation $P_m$ to have source domain label 0, while the representations $\{v3s, v4s, v5s\}_n$ and $P_n$ from image $I_n$ in RGB-D dataset to have target domain label 1. And we calculate the loss of DS by:

$$L_{DS}(DS) = L_{bce}(DS(v3s_m, v4s_m, v5s_m, P_m), 0) + L_{bce}(DS(v3s_n, v4s_n, v5s_n, P_n), 1)$$ (9)

where $L_{bce}$ 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 domain shift for each spatial position.

For $\text{DT}$, 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 assign its representations to have target domain label 0, and representations $\{v_3d, v_4d, v_5d\}_m$ and $Q_m$ from image $I_m$ to have target domain label 1. The loss for $\text{DT}$ is calculated by:

$$L_{\text{DT}}(\text{DT}) = L_{\text{bce}}(\text{DT}(v_{3t}t_n, v_{4t}t_n, v_{5t}t_n, Q_n), 0) + L_{\text{bce}}(\text{DT}(v_{3d}m, v_{4d}m, v_{5d}m, Q_m), 1)$$ (10)

To fool the saliency discriminator $\text{DS}$, our prediction model is trained to learn saliency representations $\{v_3s, v_4s, v_5s\}_t, P_t$ from $I_t$ which can be classified as source domain in $\text{DS}$. The adversarial loss for saliency branch can be calculated as:

$$L_{\text{adv},s}(E, SF, S, T) = L_{\text{bce}}(\text{DS}(v_{3s}n, v_{4s}n, v_{5s}n, P_n), 0)$$ (11)

For the depth discriminator $\text{DT}$, our prediction model is trained to learn depth representations $\{v_3d, v_4d, v_5d\}_s, Q_s$ from $I_s$ which can be classified as source domain in $\text{DT}$.

$$L_{\text{adv},d}(E, ST, S, T) = L_{\text{bce}}(\text{DT}(v_{3d}m, v_{4d}m, v_{5d}m, Q_m), 0)$$ (12)

### 3.5 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. (12)), the adversarial loss of depth branch for $I_m$ (Eq. (11)); and the losses for discriminators, which are the saliency discriminator loss (Eq. (9)), and depth discriminator loss (Eq. (10)):

$$\min_{\text{DS,DT}} L_{\text{DS}} + L_{\text{DT}}$$ (13)

$$\min_{E, ST, S, T} \lambda_s L_{\text{fin},s} + \lambda_d L_{\text{fin},d}$$

$$+ \lambda_{\text{init}} L_{\text{init},s} + \lambda_{\text{init}} L_{\text{init},d}$$

$$+ \lambda_{\text{adv},s} L_{\text{adv},s} + \lambda_{\text{adv},d} L_{\text{adv},d}$$ (14)

### 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 for our method by comparison with the state-of-art method and the ablation study.
4.1 Datasets

We evaluate our proposed method on seven widely used RGB-D datasets. NJUD [16] has 1985 images taken by a Fuji W3 stereo camera; NLPR [26] contains 1000 images constructed by Kinect; LFSD [20] has 100 images using the Lytro light field camera, STEREO [25] includes 797 images from the Internet, RGBD135 [8] contains 135 indoor images by Kinect, SIP [11] is a dataset with 929 images, which focuses on people with challenging actions. DUT-D [27] contains 1200 images with complex scenes for both indoor and outdoor situations.

For a fair comparison, we follow the common approach as in [27] as well as utilizing the DUTS [34], an RGB saliency dataset contains 10553 images for training. For DUT-D, we train our depth branch by its 800 training images and saliency branch by the DUTS training set, then evaluate the overall model on the DUT-D 400 test images. For other datasets, we use the selected 1485 NJUD images and 700 NLPR images as the RGB-D training set. These 2185 images are used to train our depth branch, and the DUTS training set are used to train our saliency branch. We then evaluate our model on the remaining images in NJUD, NLPR, and the left STEREO, LFSD, SIP, RGBD135 datasets.

4.2 Implementation Details

We apply PyTorch for our implementation using two GeForce RTX 2080 Ti GPU with 22 GB memory. For our prediction model, we use VGG19 [30] pre-trained model as backbone. And for the discriminator, we have first apply one convolution layer for each input feature/prediction and concat the latent representations. We then apply four extra convolution layers to output the one-channel classification result. Except for the last convolution layer, each convolution layer in our discriminator module is followed by a Leaky-ReLU [24] with a slope of 0.2 for negative inputs. We apply ADAM [18] optimizer for both two-stage prediction module and discriminator module, with the initial learning rate setting to 2.5e-4 and 1e-4. All the input images are resized to 256 × 256 pixels.

4.3 Evaluation Metrics

For quantitative evaluation, we adopt four widely used evaluation metrics including F-measure($F_m$) [1], mean absolute error (MAE) [3], $S$-measure($S_m$) [10] and $E$-measure($E_m$) [10]. In this paper, we report the average F-measure value as $F_m$ which is calculated by the weighted mean of the precision and recall. For MAE, it calculates the average absolute difference between the prediction and ground-truth. $S_m$ evaluates the prediction on region-aware and object-aware structural similarity. And $E_m$ captures global statistics and local pixel matching information. For MAE, the lower value indicates the method is better, while for all other metrics, the higher value indicates the method is better.
Table 1. Results on different datasets. We highlight the best two result in each column in red and blue.

| Method   | DUT-RGBD | STEREO | SIP | RGBD135 |
|----------|----------|--------|-----|---------|
|          | MAE      | F$_m$  | E$_m$ | MAE      | F$_m$  | E$_m$ | MAE      | F$_m$  | E$_m$ | MAE      | F$_m$  | E$_m$ | MAE      | F$_m$  | E$_m$ |
| DMRA     | 0.048    | 0.883  | 0.889 | 0.300   | 0.947   | 0.968 | 0.986 | 0.936   | 0.888   | 0.815 | 0.800 | 0.858 | 0.036   | 0.886  | 0.899 |
| CPFP     | 0.100    | 0.735  | 0.749 | 0.815   | 0.354   | 0.527 | 0.672 | 0.902   | 0.064   | 0.819 | 0.856 | 0.899 | 0.038   | 0.829  | 0.872 |
| 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 |
| MMCI     | 0.112    | 0.753  | 0.791 | 0.856   | 0.080   | 0.812 | 0.856 | 0.893   | 0.088   | 0.795 | 0.833 | 0.886 | 0.065   | 0.762  | 0.848 |
| PCANet   | 0.100    | 0.756  | 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 |
| 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 |
| DF       | 0.145    | 0.747  | 0.729 | 0.842   | 0.142   | 0.761 | 0.765 | 0.844   | 0.185   | 0.673 | 0.653 | 0.734 | 0.131   | 0.573  | 0.685 |
| DCMC     | 0.243    | 0.405  | 0.712 | 0.150   | 0.749   | 0.818 | 0.845 | 0.683   | 0.767   | 0.196 | 0.234 | 0.409 | 0.064   | 0.762  | 0.843 |
| CDCP     | 0.159    | 0.633  | 0.687 | 0.794   | 0.149   | 0.680 | 0.727 | 0.801   | 0.224   | 0.495 | 0.505 | 0.722 | 0.129   | 0.594  | 0.709 |
| Ours     | 0.056    | 0.886  | 0.874 | 0.928   | 0.050   | 0.858 | 0.885 | 0.934   | 0.063   | 0.859 | 0.862 | 0.914 | 0.034   | 0.872  | 0.874 |

4.4 Comparison with state-of-the-art methods

We compare our method with 9 state-of-the-art methods including 7 latest RGB-D deep learning methods: DMRA [27], CPFP [38], TANet [5], MMCI [6], PCANet [4], CTMF [13], DF [28], and 2 RGB-D traditional methods: DCMC [9], CDCP [42]. The performance of our method compared with 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 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 methods do not require access to any saliency ground-truth maps for images in RGB-D datasets, and we only use the depth data as an auxiliary task for training, which makes our model only require the RGB data at inference. By training the model in the RGB-D dataset without its saliency ground-truth maps, as well as not using any depth data in the testing process, our method still manages to be comparable with the state-of-art RGB-D saliency detection methods.

The quantitative results show that our method can achieve better results than the fully-supervised methods on some datasets like SIP and STEREO. It indicates that a larger RGB dataset can be helpful since it contains images with various situation. Even though the images from RGB dataset has considerable difference on appearance, we are able to gain useful information for helping the saliency prediction in RGB-D datasets by our cross refinement and adversarial learning method. However, the lacking of saliency ground truth still causes our method perform slightly worse on datasets such as NJUD and NLPR. The reason that we may perform worse on these two specific datasets is that all other fully-supervised methods using the saliency ground-truth maps of NJUD and NLPR.
during training. So they tend to focus on detecting saliency objects with the pattern of these two datasets during the training, which makes them perform better than our our method without using the saliency ground-truth for RGB-D images. To better demonstrate the advantage of our method, we also present some qualitative segmentation examples in Figure 5.

### 4.5 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 RGBD saliency dataset and the second stage cross-refinement prediction with prediction-guided attention module.

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

Our ablation study is evaluated on four RGB-D datasets and the result is showed in Table 3. It indicates that our baseline model provides a good initial prediction with saliency branch trained by the RGB dataset. Our feature-guided attention module helps to focus the more informative spatial positions and channels, and the FAM module enlarging the receptive field. Therefore, our method further improves performance by adding both of them.

We then add depth branch and the second-stage cross-refinement to utilize RGB-D datasets for training in the third subset model. Depth information provides spatial structure information for RGB-D images even though we do not

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

| Dataset | MAE Fm Sm Em | MAE Fm Sm Em | MAE Fm Sm Em | MAE Fm Sm Em |
|---------|--------------|--------------|--------------|--------------|
| LFSD    | 0.076 0.849 0.847 0.859 0.051 0.872 0.889 0.929 0.031 0.853 0.898 0.912 |
| NJUD    | 0.088 0.813 0.828 0.867 0.053 0.837 0.878 0.900 0.039 0.818 0.884 0.920 |
| NLPR    | 0.111 0.794 0.801 0.851 0.061 0.844 0.878 0.909 0.041 0.796 0.806 0.916 |
| MAE     | 0.132 0.779 0.787 0.840 0.079 0.813 0.859 0.882 0.059 0.730 0.565 0.872 |
| Fm      | 0.112 0.794 0.800 0.856 0.059 0.844 0.877 0.909 0.044 0.795 0.874 0.916 |
| Sm      | 0.120 0.781 0.796 0.851 0.085 0.788 0.849 0.866 0.056 0.724 0.860 0.869 |
| Em      | 0.142 0.810 0.786 0.841 0.151 0.744 0.735 0.818 0.100 0.683 0.769 0.840 |
| DMRA    | 0.088 0.813 0.828 0.867 0.053 0.837 0.878 0.900 0.039 0.818 0.884 0.920 |
| CPFP    | 0.120 0.781 0.796 0.851 0.085 0.788 0.849 0.866 0.056 0.724 0.860 0.869 |
| TANet   | 0.112 0.794 0.800 0.856 0.059 0.844 0.877 0.909 0.044 0.795 0.874 0.916 |
| CMF     | 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 |
| DCPC    | 0.155 0.815 0.754 0.842 0.167 0.715 0.763 0.796 0.196 0.328 0.556 0.685 |
| Ours    | 0.102 0.830 0.819 0.878 0.084 0.809 0.860 0.907 0.048 0.806 0.863 0.904 |

### Table 3. Ablation Study on our proposed method. We highlight the best result in each column in red.
have access to their saliency ground-truth maps. Normally, by adding depth branch and our second stage for cross refinement, it gains improvements for saliency detection on RGB-D images. However, only adding depth branch may cause a negative effect since the saliency branch is supervised by the RGB dataset only, and depth branch is supervised by the RGB-D dataset only. The difference between two kinds of datasets on appearance and distribution affect the generalization of the learned model. The comparison of the third and fourth subset model verify the effectiveness of our adversarial learning by adding the discriminator module. It aligns the representation on two datasets for each branch, and tackles the domain shift on different datasets.

5 Conclusions

In this paper, we propose a novel synergistic saliency and depth prediction method for RGB-D saliency detection 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-labelled RGB saliency datasets, avoid using saliency ground-truth maps from RGB-D datasets during training, and require only RGB 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, allowing two branches to provide complementary information for each other; a discriminator with adversarial learning to reduce the impact of the shift caused by difference on distributions of different datasets. Evaluation on seven RGB-D datasets demonstrates the effectiveness of our method, by achieving a comparable method with the state-of-art fully-supervised RGB-D saliency methods.
References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1597–1604. IEEE (2009)
2. Bak, S., Carr, P., Lalonde, J.F.: Domain adaptation through synthesis for unsupervised person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 189–205 (2018)
3. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE transactions on image processing 24(12), 5706–5722 (2015)
4. 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. pp. 3051–3060 (2018)
5. Chen, H., Li, Y.: Three-stream attention-aware network for rgb-d salient object detection. IEEE Transactions on Image Processing 28(6), 2825–2835 (2019)
6. 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, 376–385 (2019)
7. 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. pp. 3339–3348 (2018)
8. 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. pp. 23–27 (2014)
9. 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(6), 819–823 (2016)
10. 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. pp. 4548–4557 (2017)
11. 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)
12. 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. pp. 2672–2680 (2014)
13. 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(11), 3171–3183 (2017)
14. 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. pp. 3005–3012. IEEE (2012)
15. 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. pp. 1498–1507 (2018)
16. 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). pp. 1115–1119. IEEE (2014)
17. Kim, S., Choi, J., Kim, T., Kim, C.: Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6092–6101 (2019)

18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)

19. Lee, H., Kim, D.: Salient region-based online object tracking. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1170–1177. IEEE (2018)

20. 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. pp. 2806–2813 (2014)

21. 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. pp. 3917–3926 (2019)

22. 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. pp. 6778–6787 (2019)

23. 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. pp. 2507–2516 (2019)

24. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: in ICML Workshop on Deep Learning for Audio, Speech and Language Processing. Citeseer (2013)

25. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 454–461. IEEE (2012)

26. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: a benchmark and algorithms. In: European conference on computer vision. pp. 92–109. Springer (2014)

27. 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. pp. 7254–7263 (2019)

28. 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(5), 2274–2285 (2017)

29. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)

31. Tsai, Y.H., Hung, W.C., Schultet, 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. pp. 7472–7481 (2018)

32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)

33. 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. pp. 2517–2526 (2019)

34. 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. pp. 136–145 (2017)

35. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)

36. 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. pp. 2048–2057 (2015)

37. 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. pp. 5031–5040 (2019)

38. 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. pp. 3927–3936 (2019)

39. 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). pp. 199–204. IEEE (2019)

40. 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. pp. 3008–3014 (2017)

41. 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. pp. 14–23. Springer (2017)

42. 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. pp. 1509–1515 (2017)