# Multi-spectral Salient Object Detection by Adversarial Domain Adaptation

**Shaoyue Song,[1*] Hongkai Yu,[2*] Zhenjiang Miao,[1*] Jianwu Fang,[3] Kang Zheng,[4] Cong Ma,[1] Song Wang[4,5]**

[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[2]Department of Computer Science, University of Texas-Rio Grande Valley, Edinburg, TX, USA
[3]School of Electronic and Control Engineering, Chang'an University, Xi'an, China
[4]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA
[5]School of Computer Science and Technology, Tianjin University, Tianjin, China

## Abstract

Although there are many existing research works about the salient object detection (SOD) in RGB images, there are still many complex situations that regular RGB images cannot provide enough cues for the accurate SOD, such as the shadow effect, similar appearance between background and foreground, strong or insufficient illumination, etc. Because of the success of near-infrared spectrum in many computer vision tasks, we explore the multi-spectral SOD in the synchronized RGB images and near-infrared (NIR) images for the both simple and complex situations. We assume that the RGB SOD in the existing RGB image datasets could provide references for the multi-spectral SOD problem. In this paper, we first collect and will publicize a large multi-spectral dataset including 780 synchronized RGB and NIR image pairs for the multi-spectral SOD problem in the simple and complex situations. We model this research problem as an adversarial domain adaptation from the existing RGB image dataset (source domain) to the collected multi-spectral dataset (target domain). Experimental results show the effectiveness and accuracy of the proposed adversarial domain adaptation for the multi-spectral SOD.

## Introduction

In computer vision, salient object detection (SOD) that aims at finding out the salient objects in a given image is helpful to discover the objects and well understand the image scene, so the SOD techniques could benefit many applications, such as image scene understanding (Zhang, Du, and Zhang 2014), image segmentation (Wang et al. 2018), object tracking (Zhang et al. 2018), common object discovery (Yu et al. 2018), etc. There are many existing research works about the SOD in RGB images such as (Cheng et al. 2015; Chen et al. 2018), which have achieved advanced performance in regular simple situations. However, there are still many complex situations that regular RGB images cannot provide enough cues for the accurate SOD, such as the shadow effect, similar appearance between background and foreground, strong or insufficient illumination, as shown in Fig. 1.

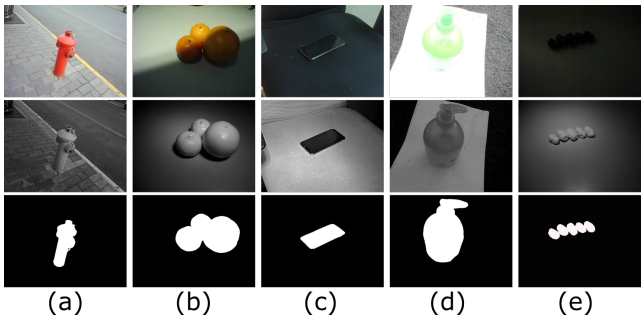


Figure 1: Sample images from the collected multi-spectral dataset in simple and complex situations: (a) simple/normal situation, (b) shadow effect, (c) similar appearance between background and foreground, (d) strong illumination, and (e) insufficient illumination. From top to the bottom: RGB image, synchronized NIR image, annotated ground truth for the SOD.

Recently, the near-infrared spectrum has shown successes in many computer vision tasks. Near-infrared (NIR) image is one of the image modalities often used to help the RGB image tasks such as the low-light image enhancement, image restoration, image dehazing, robust scene category recognition (Brown and Süsstrunk 2011), face recognition robust to illumination variations (He et al. 2018), image quality and context improvement to the changeable weather (Jiang et al. 2019), etc. For example, as shown in Fig. 1, the RGB images might show low discriminative contrast in complex situations, while the synchronized NIR images might display a better contrast to human beings. In many real-world applications like robotics, autonomous vehicles, and video surveillance, the multi-spectral images including RGB and NIR images are available, so it is highly desired to systematically study the multi-spectral SOD problem. Therefore, in this paper, we explore the multi-spectral SOD problem in the synchronized RGB and NIR images for the both simple and complex situations.

Different from many SOD methods extracting effective feature representations for saliency detection, we assume that the RGB SOD in the existing public RGB image

*Co-corresponding authors.

datasets (such as the well-labelled MSRA-B (Liu et al. 2010), DUTS (Wang et al. 2017b), HKU-IS (Li and Yu 2015)) could provide references for the multi-spectral SOD problem. Existing datasets using NIR images (Wang, Zhu, and Yuan 2013; Wang et al. 2013) for SOD are very small with only dozens of RGB-NIR image pairs. In this paper, we first collect and will publicize a large multi-spectral dataset including 780 synchronized RGB and NIR image pairs for the multi-spectral SOD problem in the simple and complex situations. We model this research problem as an adversarial domain adaptation from the existing RGB image dataset (source domain) to the collected multi-spectral dataset (target domain). The main contributions of this paper are as follows:

1. To the best of our knowledge, this is the first work to systematically study the research problem of multi-spectral salient object detection using the synchronized RGB and NIR images for the both simple and complex situations.
2. we first collect a large multi-spectral dataset of 780 synchronized RGB and NIR image pairs including simple and complex situations for the SOD problem. Each image pair has been carefully annotated with the pixel-level SOD ground truth. .
3. We propose a new method for the multi-spectral SOD based on the adversarial domain adaptation from the existing RGB image dataset (source domain) to the collected multi-spectral dataset (target domain).

## Related Work

**RGB image SOD:** Salient object detection is to find the visual salient object/region which mostly attracts human attention in a given image. It is a fundamental task in computer vision. The traditional methods like RC (Cheng et al. 2015), LRK (Shen and Wu 2012), CWS (Fu, Cao, and Tu 2013), FT (Achanta et al. 2009) usually concentrate on some specific low-level features and certain prior information like connectivity prior (Vicente, Kolmogorov, and Rother 2008), background prior (Wei et al. 2012). Recently, by the powerful representation of deep learning based methods, the SOD task performance is improved a lot. In deep learning based methods (Li and Yu 2016; Chen et al. 2018; Wu, Su, and Huang 2019), sufficient training data is important. There are many well-labelled datasets of RGB images such as MSRA-B (Liu et al. 2010), DUTS (Wang et al. 2017b), HKU-IS (Li and Yu 2015), etc. In this paper, we suppose that the RGB SOD datasets can provide references and guides for the multi-spectral SOD problem.

**Mutli-spectral SOD and related datasets:** The most related datasets to our SOD problem is the existing multi-spectral datasets including NIR images (Wang, Zhu, and Yuan 2013; Wang et al. 2013). They collect several RGB-NIR image pairs to explore the near-infrared clues in the saliency detection. However, their datasets only have a small number of image pairs. Some researches also concentrate on other modalities of images to help the SOD with regular RGB images. The depth image is considered to explore the RGB-Depth SOD (Qu et al. 2017). The thermal infrared dataset is also collected for SOD (Tu et al. 2019). For the SOD tasks, it is desirable to include more multi-spectral cues and models to improve the SOD performance.

**Adversarial Domain Adaptation:** In the research of adversarial domain adaptation, generative adversarial learning (Goodfellow et al. 2014) could be used to reduce the domain shifts across different domains. Typically, a generator and a discriminator are trained against each other (Tzeng et al. 2017; Vu et al. 2019). The generator is trained to confuse the discriminator, while the discriminator is trained to classify the features coming from different domains. Following this procedure, the domain bias could be reduced leading to the improved performance (Tzeng et al. 2017; Vu et al. 2019; Benjdira et al. 2019).

## Multi-spectral SOD Dataset

We collect a new dataset consisting of 780 RGB-NIR image pairs of the same scene in this paper. The image pairs mainly contain some ordinary objects in the indoor scene (409 image pairs) and outdoor scene (371 image pairs).

**Dataset statistics.** Since the research target of this paper is to explore the multi-spectral SOD problem in both simple/normal and complex situations, we collect the RGB-NIR image pairs in both simple/normal and complex situations. For the normal situation, we consider the salient objects in the normal indoor and outdoor environments. For the complex situations, we collect the images of salient objects in the challenging light illumination (213 image pairs), shadow influence (165 image pairs), and similar appearance of background and foreground (169 image pairs). The data distribution of the collected multi-spectral SOD Dataset is shown in Fig. 2 (a) and (b).

The collected RGB image and the corresponding NIR image are synchronized and aligned towards the same salient object(s), as shown in Fig. 1. The original NIR image is an image of single channel, then we duplicate it to be a three-channel image same as that of the RGB image.

**Image capture and annotation.** We capture the multi-source image data by a multi-spectral camera developed by ourselves with an estimate cost of 100 to 200 dollars. The sensor simultaneously captures RGB and near-infrared bands with two separate lens. In order to make the details in near-infrared band clear, we equipped the near-infrared supplemental lamp, where the wavelength of the near-infrared band is 850 nm. The multi-spectral camera could capture the synchronized and aligned RGB-NIR image pairs. Each image size is $640 \times 480$ pixels.

We carefully annotate each image pair with the help of 5 computer vision researchers who have clearly learned how to define the salient object(s) in a given RGB-NIR image pair. The participants are asked to manually label the salient object(s) by pixel-level annotations. The average ground-truth distribution of proposed multi-spectral SOD dataset is shown in Fig. 2 (c).

## Methodology

In this section, we will firstly introduce the proposed unsupervised adversarial domain adaptation for the multi-spectral SOD problem, i.e., training on the existing RGB SOD dataset (source domain) and testing on the proposed multi-spectral SOD dataset (target domain). In addition, we further introduce the supervised domain adaptation for the multi-spectral SOD problem.

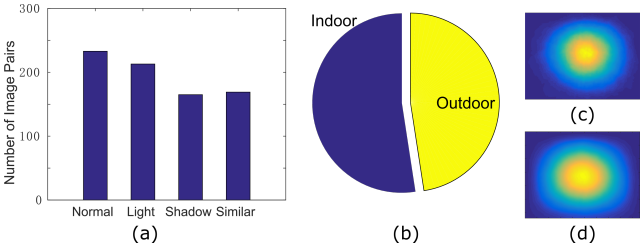


Figure 2: Statistics for the proposed multi-spectral SOD dataset: (a) distribution under simple/normal and complex situations, (b) distribution under indoor and outdoor scenes, (c) average ground-truth on the proposed dataset. (d) average ground-truth on MSRA-B dataset (Liu et al. 2010).

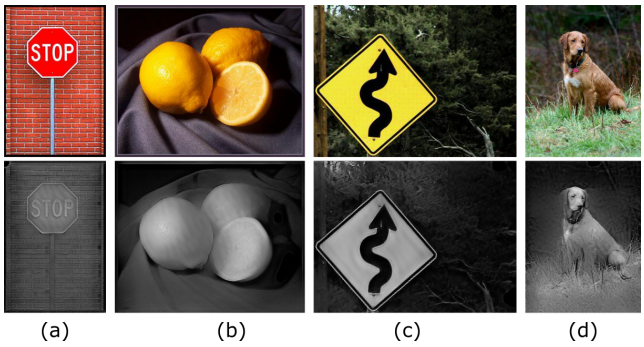


Figure 3: Examples of pseudo-NIR image generation. Top row: RGB images and bottom row: generated pseudo-NIR images. The CycleGAN model is trained between the RGB images of the MSRA-B dataset and the NIR images of the proposed multi-spectral SOD dataset.

In the unsupervised scenario, we assume the source-domain RGB images $X_S^{rgb}$ and their pixelwise SOD ground-truth labels $y_s$ are drawn from the source domain distribution $S$, and the target-domain image pairs $X_T^{rgb}$ and $X_T^{nir}$ without the pixelwise SOD ground-truth label are drawn from a target domain distribution $T$. The goal of the proposed method is to learn the SOD model $G(\cdot)$ under the supervision of $S$ and perform well on the test images of $T$. A domain classifier $D(\cdot)$ is defined to reduce the domain shift between $S$ and $T$ with the domain label $l$, where the domain label only indicates the images coming from $S$ or $T$. The whole framework of the proposed method is shown in Fig. 4.

### CycleGAN based pseudo-NIR image generation

One challenge in the domain adaptation problem discussed in this paper is that all the existing RGB image datasets for SOD only contain RGB color images and do not have the corresponding NIR images. This challenge will affect the performance due to the lack of NIR information in the source domain $S$. In order to solve this problem, we employ an image-to-image translation to synthesize the pseudo NIR images for the source domain $S$. Because we do not have the paired RGB-NIR image data for the existing RGB image datasets like MSRA-B (Wang et al. 2017a), this translation is an unpaired image-to-image transfer, which can be achieved by the advanced CycleGAN (Zhu et al. 2017).

CycleGAN (Zhu et al. 2017) is a popular unpaired image-to-image translation framework to learn the mapping between two domains with unpaired images, where the transferred images from $S$ could be similar to the expected image styles in the target domain $T$. Given the source-domain RGB images $X_S^{rgb}$ and target-domain NIR images $X_T^{nir}$ of the proposed multi-spectral SOD dataset, following the network structure and setup in CycleGAN (Zhu et al. 2017), we can learn a generator $G_{ST}$, which represents the mapping: $X_S^{rgb} \rightarrow X_T^{nir}$. In our experiments, the trained $G_{ST}$ is used to generate pseudo-NIR images $X_S^{nir}$ for each RGB image of the source domain $S$. The examples of the CycleGAN based pseudo-NIR image generation are shown in Fig. 3. With the help of CycleGAN based pseudo-NIR image generation, the cross-domain data distribution discrepancy is somewhat reduced. Experimental results also display the effectiveness of the pseudo-NIR image generation in domain shift reduction.

### Two-branch SOD network

With the help of the generated pseudo-NIR images, both the source domain $S$ and target domain $T$ have synchronized RGB-NIR image pairs. In order to fully use the NIR spectrum image to enhance the SOD task, we propose a two-branch SOD network for the multi-spectral SOD. The two-branch SOD network has paired images as the input, and outputs the corresponding saliency map. We adopt an original Fully Convolutional Networks (FCN) (Long, Shelhamer, and Darrell 2015) with two branches to output the saliency prediction. FCN is widely used in the saliency detection to predict the probability of each pixel as the salient objects (Li and Yu 2016).

As shown in Fig. 5, the proposed SOD network $G(\cdot)$ has two branches: the RGB FCN branch $G_{rgb}(\cdot)$ taking RGB image as the input and the NIR FCN branch $G_{nir}(\cdot)$ taking NIR image as the input. The two branches are with shared weights that can be trained in an end-to-end way. For each branch, we modify the original FCN to output a two-channel map by applying the softmax function on each pixel, i.e., obtaining the probability to be foreground or background for each pixel. $G(\cdot) = G_{rgb}(\cdot) \bigoplus G_{nir}(\cdot)$, where $\bigoplus$ means pixel-wise addition. In our experiment, we adopt VGG16 (Simonyan and Zisserman 2014) as our backbone network for FCN, and other FCN models can also be applied to our proposed framework. The proposed two-branch

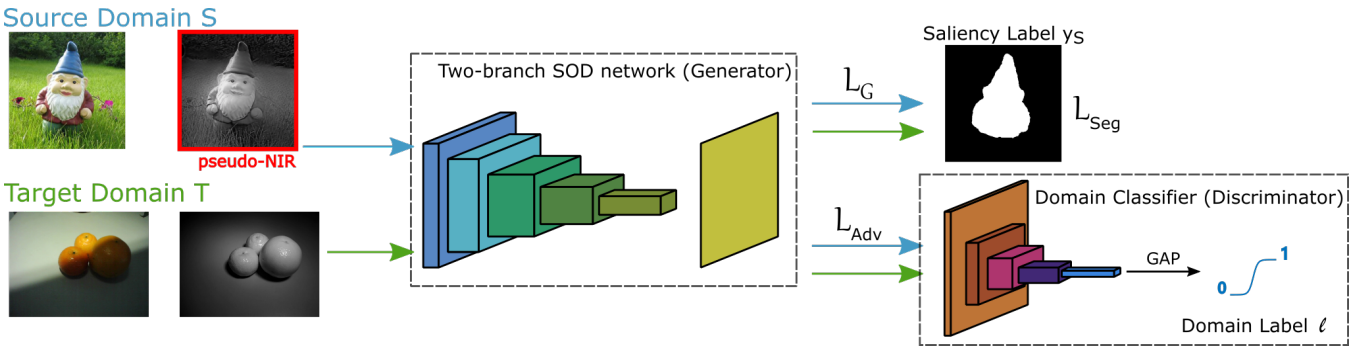


Figure 4: Framework of the proposed adversarial domain adaptation method for the unsupervised multi-spectral SOD. It consists of a two-branch SOD network (Generator) and a domain classifier (Discriminator). The source domain $S$ is an existing RGB SOD dataset like MSRA-B (Wang et al. 2017a) with the pixelwise ground-truth labels and the target domain $T$ is the proposed multi-spectral SOD dataset without the pixelwise ground-truth labels.

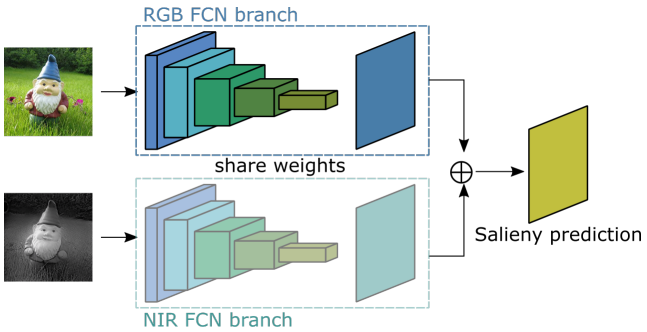


Figure 5: The proposed two-branch SOD network with the RGB and NIR (or pseudo-NIR) image as the inputs.

SOD network is simple but efficient to capture the united multi-spectral cues from RGB and NIR images.

## Unsupervised adversarial domain adaptation

Directly applying the two-branch SOD network trained on the source domain $S$ to test the images on the target domain $T$ might only obtain low performance due to the domain distribution discrepancy. With the assumption that the source domain $S$ could provide references to the target domain $T$, we think the two domains $S$ and $T$ have some latent feature spaces that are domain-invariant for the multi-spectral SOD problem. It is hard to directly find the shared latent feature space, thus we use adversarial learning for this task. In particular, we treat the proposed two-branch SOD network as a Generator and then we apply a domain classifier as the Discriminator as defined in Fig. 4. By adversarial learning, the Generator learns to generate the SOD map to fool the Discriminator, while the Discriminator will learn to classify the image pair coming from $S$ or $T$. In this adversarial way, the Generator finally learn a network to generate the multi-spectral SOD map that cannot be classified by the domain classifier, which means that we find a network to extract the domain-invariant features.

The domain classifier network $D(\cdot)$ used in the proposed method is built as a discriminator network by following the Discriminator in the DCGAN (Radford, Metz, and Chintala 2015) as a reference. $D$ has five stacked strided convolutional layers with $3 \times 3$ kernel and numbers of channels as $\{64, 128, 256, 512, 1\}$. The stride is setting up as stride = 2 except the last convolutional layer. The model of the discriminator network is much smaller than the generator network. LeakyReLU activation layer is followed with convolutional layers except for the last layer. As mentioned in (Radford, Metz, and Chintala 2015), using strided convolution allows the network to learn its own spatial down pooling and using leakyReLU activation works well for higher resolution modeling. The Global Average Pooling (GAP) and Sigmoid activation function are applied to output the domain label prediction (1 for domain $S$ and 0 for domain $T$). In our proposed framework, we take the two-branch SOD network $G(\cdot)$ as a domain feature generator which is optimized by minimizing a standard supervised pixel-wise cross entropy loss $\mathcal{L}_{Seg}$:

$$\mathcal{L}_{Seg} = -\sum_{X_S}[y_s \log(G(X_S)) + (1 - y_s)\log(1 - G(X_S))], \tag{1}$$

where $X_S$ is a source-domain RGB-NIR image pair, $G(X_S)$ is the predicted saliency map. $y_S$ is the two-class pixel-wise ground-truth map of salient and non-salient classes. Like (Goodfellow et al. 2014; Vu et al. 2019), the domain classifier $D$ is trained to discriminate $G(X)$ coming from the source or target domains, and at the same time, the two-branch SOD model $G(\cdot)$ as the generator is trained to confuse the discriminator $D$. Suppose $\mathcal{L}_D$ denotes the cross entropy domain classification loss and $F = D(G(\cdot))$, and we define the domain label $l_s = 1$ for the image pair from the source domain and $l_t = 0$ for the image pair from the target domain, and then the adversarial loss for the domain classifier $D$ is:

$$\mathcal{L}_{Adv} = \sum_{X_S} \mathcal{L}_D(F(X_S), l_s) + \sum_{X_T} \mathcal{L}_D(F(X_T), l_t). \tag{2}$$

The loss for training $G(\cdot)$ is defined as combing Eq. (1) and Eq. (2) as:

$$\mathcal{L}_G = \sum_{X_S} \mathcal{L}_{Seg}(X_S, y_s) + \lambda_1 \sum_{X_S} \mathcal{L}_D(F(X_S), l_t) + \lambda_2 \sum_{X_T} \mathcal{L}_D(F(X_T), l_s), \quad (3)$$

where $\lambda_1$ and $\lambda_2$ are the balance weights and we set them as 1 in our experiments. The learning can be summarized as the following optimization problem:

$$\min_{\theta_G} \mathcal{L}_G, \quad (4)$$

$$\min_{\theta_D} \mathcal{L}_{Adv}. \quad (5)$$

During the training procedure, we alternately optimize the network parameters $\theta_G$ for $G(\cdot)$ by optimizing Eq. (4) and the network parameters $\theta_D$ for $D(\cdot)$ by optimizing Eq. (5).

### Supervised domain adaptation via fine-tuning

Besides the unsupervised scenario, we also consider the supervised domain adaptation for the multi-spectral SOD task via fine-tuning. For the supervised scenario, we split the collected multi-spectral SOD dataset into training, validation and testing subsets as the ratio of 5:1:4 same as the split principle in the MSRA-B dataset (Jiang et al. 2013). The image pairs are randomly selected from the simple and different complex situations following the split ratio. Given a pre-trained model, it can be fine-tuned on the training and validation subsets of the collected multi-spectral SOD dataset for a supervised domain adaptation.

The two-branch SOD network $G(\cdot)$ is adopted for the supervised task. We mainly consider this supervised learning problem as a domain adaptation from the pre-trained models on some existing dataset to the collected multi-spectral SOD dataset. We study the three fine-tuning strategies with different initialized pre-trained models using the ImageNet dataset, the existing RGB SOD dataset, and the proposed unsupervised domain adaptation model.

## Experiment

### Source-domain dataset and the pseudo-NIR images

For unsupervised domain adaptation SOD task, we choose the MSRA-B dataset (Jiang et al. 2013; Wang et al. 2017a) as our source domain, and our multi-spectral SOD dataset (780 RGB-NIR image pairs) is taken as the target domain. MSRA-B dataset includes 5000 RGB images which contains various image contents of natural scenes, animals, planets, etc. The dataset is divided into three parts as the ratio of 5:1:4 (training: 2500 images, validation: 500 images, testing: 2000 images) (Jiang et al. 2013). For training the CycleGAN model, we take all the 2500 training images in MSRA-B and all the 780 NIR images in multi-spectral SOD dataset as the two domains for image translation.

For training the CycleGAN model, we choose the cross-entropy loss mode, image buffer is set as 50 inspired by (Zhu et al. 2017) and other hyper-parameters like input image size as 256, are following the default setup in their public code. To balance the training time and image quality, we keep the 50th training epoch model as our generator to synthesize the pseudo-NIR images for MSRA-B dataset. Some typical images of the original and synthetic image pairs are shown in Fig. 3, and we can see that the generated pseudo-NIR images are reasonable and similar as the real NIR images.

### Evaluation Metrics

We evaluate the proposed multi-spectral SOD method performance using Precision-Recall (PR) curve, maximum F-measure (max-F), and Mean Absolute Error (MAE). We also evaluate the average precision, recall and F-measure with an adaptive threshold that is twice the mean value of the saliency map (Yu et al. 2018). The value of F-measure is defined as $F_\gamma = \frac{(1+\gamma^2)\times Precision \times Recall}{\gamma^2 \times Precision + Recall}$, where $\gamma^2$ is set to 0.3 as suggested in (Li and Yu 2016). When given a threshold $\theta (\theta \in [0, 1])$ to a saliency map, we can get a binary mask of it. Then the precision and recall can be computed by comparing the generated binary saliency mask and the ground truth. The PR curve is obtained by continuously varying $\theta$. The PR curve of a dataset is computed from the average precision and recall value over the whole dataset. The MAE error (Perazzi et al. 2012) is calculated as the average absolute pixel-wise difference between the predicted saliency map and the binary ground truth.

Table 1: Performance comparisons to other methods on the collected multi-spectral SOD dataset.

| Method | Subset | Precision | Recall | Fmeasure | MAE | maxF |
|---|---|---|---|---|---|---|
| RC | RGB | 0.6612 | 0.6812 | 0.6310 | 0.1621 | 0.7032 |
| | RGBN | 0.6743 | 0.7785 | 0.6641 | 0.1455 | 0.7333 |
| LRK | RGB | 0.5865 | 0.6923 | 0.5640 | 0.1743 | 0.6588 |
| | RGBN | 0.5892 | 0.7018 | 0.5658 | 0.1786 | 0.6640 |
| CWS | RGB | 0.5916 | 0.5625 | 0.5471 | 0.2428 | 0.6137 |
| | RGBN | 0.5723 | 0.5225 | 0.5059 | 0.2452 | 0.5784 |
| FT | RGB | 0.3496 | 0.3954 | 0.3322 | 0.1974 | 0.3622 |
| | RGBN | 0.3952 | 0.4320 | 0.3701 | 0.1934 | 0.4216 |
| DCL | RGB | 0.7789 | 0.7636 | 0.7461 | 0.0738 | 0.7885 |
| | RGBN | 0.7907 | **0.8436** | 0.7791 | 0.0768 | 0.8367 |
| SOD16s$^+$ | RGB$^2$ | 0.6946 | 0.7770 | 0.6806 | 0.0851 | 0.7502 |
| | RGBN | 0.7209 | 0.8259 | 0.7137 | 0.0764 | 0.8093 |
| SOD8s$^+$ | RGB$^2$ | 0.7907 | 0.7666 | 0.7572 | 0.0692 | 0.7966 |
| | RGBN | **0.8266** | 0.8207 | **0.8030** | **0.0611** | **0.8458** |

### Unsupervised domain adaptation based SOD task

In the unsupervised adversarial domain adaptation task, we aim at training a SOD model with the existing RGB image dataset with annotation to perform well on our multi-spectral SOD dataset without annotation. We implement our networks using PyTorch running on a single Tesla P40 GPU. The FCN8s network (Long, Shelhamer, and Darrell 2015) using VGG16 is used as our backbone model, and we also conduct experiments on the FCN16s network using VGG16. During the training procedure, we set the batch size as 1.

In the unsupervised situation, our proposed method is denoted as "SOD*$^+$", where "SOD" means the proposed two-branch FCN network (Generator), and "*" means the FCN

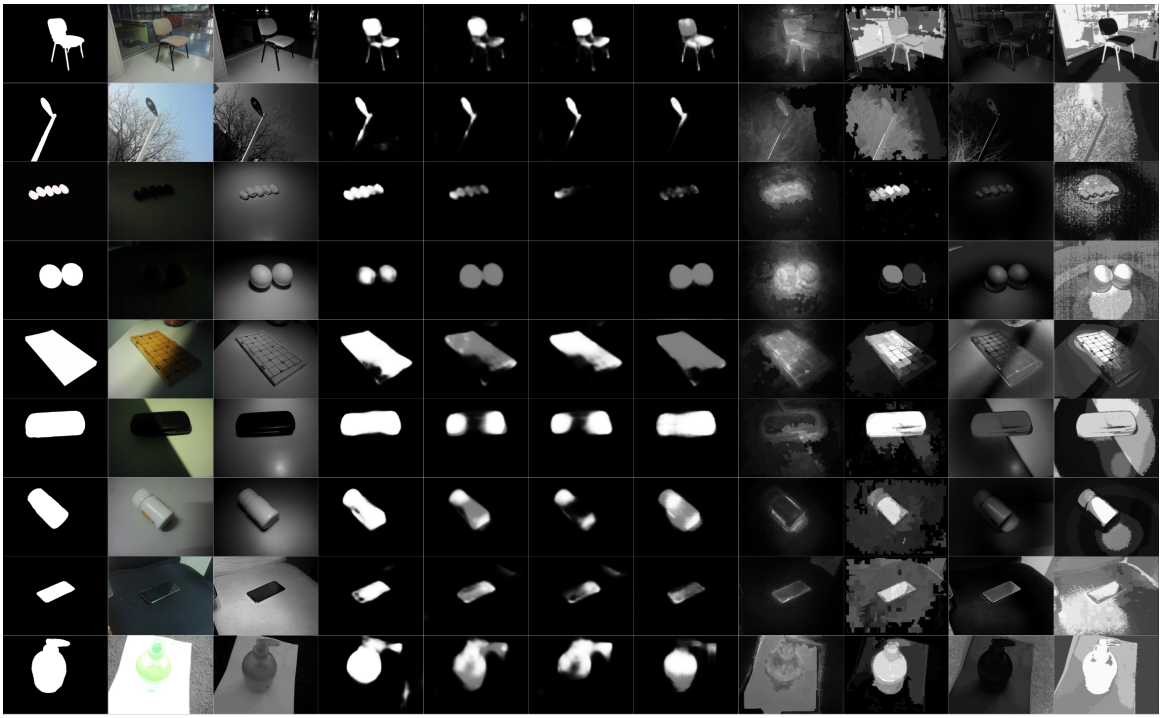


Figure 6: Sample results of unsupervised multi-spectral salient object detection: (a) ground-truth, (b) RGB image, (c) NIR image, (d) RGBN results on SOD8s$^+$ (the proposed method with unsupervised domain adaptation), (e-f) RGBN and RGB results on FCN8s, (g) DCL, (h) LRK, (i) RC, (j) FT, and (k) CWS.

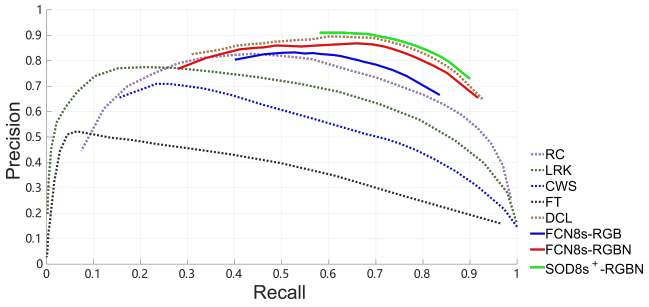


Figure 7: PR curves of different unsupervised SOD methods on the proposed multi-spectral dataset.

backbone for one branch, and "+" indicates the proposed adversarial domain adaptation method. "SOD*" is the models trained with the two-branch SOD network without the proposed adversarial domain adaptation method. "FCN*" specifies training the models with the original single branch FCN network and then testing it for RGB and NIR images independently and then merging the results. In unsupervised SOD task, both of the basic FCN model and the two-branch FCN model are initialized by the pre-trained VGG16 model on ImageNet (Russakovsky et al. 2015). In our experiments, names "*8s" or "*16s" indicate the backbone network as FCN8s or FCN16s using VGG16, respectively.

We use the training and validation set of MSRA-B dataset for training and validation. The image pairs in the collected multi-spectral SOD dataset are treated as the testing set. Firstly, we train an original single branch FCN model on source domain $S$ as our baseline model, indicated as "FCN*". All the images in our proposed dataset is tested on the well-trained FCN model. Then the two-branch SOD network "SOD*" is trained using both of the original RGB and the pseudo-NIR images of the MSRA-B dataset. The stochastic gradient descent optimizer is adopted for training. We set the momentum as $0.99$, weight decay as $0.0005$. As for learning rate, we follow the setup in (Wada 2017), i.e., lr = $10^{-10}$ for those layers with bias = False, and $2\times$ lr for the layers with bias = True. Finally, the proposed domain adaptation based model "SOD*$^+$" is trained with the initial parameters of the trained two-branch FCN model "SOD*". During training procedure of domain classifier (Discriminator), an ADAM optimizer is adopted, and the initial learning rate is $1 \times 10^{-4}$.

During testing, except for the "SOD*" models that can provide results of paired multi-spectral images simultaneously, other comparison methods can only provide saliency prediction for RGB images. For this kind of methods, we just treat both the RGB and NIR image as separate inputs. Pixelwise average results of the corresponding RGB and NIR saliency maps are used to merge the image pair's results. The results merged the RGB and NIR information are specified as "RGBN" in the tables of this paper. "RGB" in-

Table 2: Performance on the unsupervised domain adaptation for the multi-spectral SOD on the collected dataset.

| Method | Subset | Precision | Recall | Fmeasure | MAE | maxF |
|---|---|---|---|---|---|---|
| FCN16s | RGB | 0.7096 | 0.7262 | 0.6773 | 0.0948 | 0.7106 |
| | RGBN | 0.7201 | 0.7986 | 0.7079 | 0.0984 | 0.7632 |
| SOD16s | $RGB^2$ | 0.6007 | 0.7700 | 0.5998 | 0.1135 | 0.7032 |
| | RGBN | 0.6380 | **0.8367** | 0.6444 | 0.0996 | 0.7773 |
| $SOD16s^+$ | $RGB^2$ | 0.6946 | 0.7770 | 0.6806 | 0.0851 | 0.7502 |
| | RGBN | 0.7209 | 0.8259 | 0.7137 | 0.0764 | 0.8093 |
| FCN8s | RGB | 0.7737 | 0.7380 | 0.7321 | 0.0801 | 0.7650 |
| | RGBN | 0.7875 | 0.8175 | 0.7695 | 0.0829 | 0.8194 |
| SOD8s | $RGB^2$ | 0.7613 | 0.7575 | 0.7320 | 0.0766 | 0.7688 |
| | RGBN | 0.8170 | 0.8117 | 0.7911 | 0.0700 | 0.8200 |
| $SOD8s^+$ | $RGB^2$ | 0.7907 | 0.7666 | 0.7572 | 0.0692 | 0.7966 |
| | RGBN | **0.8266** | 0.8207 | **0.8030** | **0.0611** | **0.8458** |

Table 3: Performance on the supervised domain adaptation for the multi-spectral SOD on the collected dataset.

| Method | Subset | Precision | Recall | Fmeasure | MAE | maxF |
|---|---|---|---|---|---|---|
| VGG16s | $RGB^2$ | 0.6976 | 0.8413 | 0.7016 | 0.0799 | 0.7705 |
| | RGBN | 0.7689 | 0.8783 | 0.7728 | 0.0653 | 0.8303 |
| SOD16s | $RGB^2$ | 0.7101 | 0.8654 | 0.7196 | 0.0716 | 0.8024 |
| | RGBN | 0.7742 | 0.8970 | 0.7829 | 0.0570 | 0.8618 |
| $SOD16s^+$ | $RGB^2$ | 0.6807 | 0.8855 | 0.6991 | 0.0714 | 0.8075 |
| | RGBN | 0.7385 | 0.9137 | 0.7559 | 0.0579 | 0.8661 |
| VGG8s | $RGB^2$ | 0.7466 | 0.9083 | 0.7630 | 0.0586 | 0.8309 |
| | RGBN | 0.7945 | **0.9270** | 0.8089 | 0.0464 | 0.8793 |
| SOD8s | $RGB^2$ | 0.7588 | 0.9022 | 0.7702 | 0.0548 | 0.8433 |
| | RGBN | 0.8117 | 0.9238 | 0.8276 | 0.0421 | 0.8904 |
| $SOD8s^+$ | $RGB^2$ | 0.7955 | 0.8904 | 0.8010 | 0.0498 | 0.8543 |
| | RGBN | **0.8502** | 0.9156 | **0.8533** | **0.0389** | **0.9031** |

dicates the results by only testing on the RGB images. For the two-branch models, "$RGB^2$" indicates the input of each branch in the proposed two-branch SOD network is the same RGB image during testing. The experimental results about the unsupervised domain adaptation are summarized in Table 1 and Table 2.

Table 1 shows the SOD results compared with other methods. We compare our unsupervised method with some salient object detection methods as RC (Cheng et al. 2015), LRK (Shen and Wu 2012), CWS (Fu, Cao, and Tu 2013), FT (Achanta et al. 2009), and DCL (Li and Yu 2016). The first four methods are feature-based traditional methods and the last one is deep learning based method. Figure 6 shows sample results of different SOD methods and the PR curve of the related results are also shown in Fig. 7. From the results, we can find that the proposed method performs better than the other SOD methods on Precision, F-measure, MAE, and maxF metrics. Table 2 shows the performance change of each component of the proposed method. Taking FCN8s as an example, adding the synthetic pseudo-NIR images for training by "SOD8s" will get better results than FCN8s, and then further adding the proposed adversarial domain adaptation by "$SOD8s^+$" will obtain improved results. The same change trend happens to the proposed method using FCN16s as a baseline. In addition, both Table 1 and Table 2 show that using RGB-NIR image pairs together could achieve better results than only using RGB images for the saliency detection, especially in images under complex situation.

### Supervised domain adaptation based SOD task

We also evaluate the supervised domain adaptation on the collected multi-spectral SOD dataset. We mainly consider the following three initializations for fine-tuning:

1. "VGG": initializing the network $G(\cdot)$ with a pre-trained VGG16 model on the ImageNet dataset.
2. "SOD*": initializing the network with the parameter of the pre-trained model "SOD*" (trained on MSRA-B dataset without the proposed adversarial domain adaptation).
3. "$SOD*^+$": initializing the network with the parameter of the trained model "$SOD*^+$"(trained on MSRA-B dataset with the proposed adversarial domain adaption).

Table 3 shows the results of different initializations for fine-tuning. We see that the network initialized with a higher performance on the unsupervised task can help to learn a supervised model with a better performance. For example, using the pre-trained model by "$SOD8s^+$" gets the best performance, i.e., maxF=0.9031, using RGB and NIR image pairs together. We can see that using MSRA-B dataset for training by "SOD8s" could provide a better results than directly using the pre-trained model on ImageNet. The domain adaption can also be realized by fine-tuning the pre-trained model.

## Conclusion

In this paper, we systematically studied the multi-spectral salient object detection problem. We first proposed a new large dataset including 780 synchronized image pairs in both simple and complex situations and their pixelwise ground truth for this research problem. Different with traditional saliency detection methods, in this paper, we proposed a new adversarial domain adaptation method for the multi-spectral salient object detection by making better usage of the existing RGB saliency detection dataset.

The experimental results including unsupervised and supervised settings show that the multi-spectral images could better detect the salient objects than single RGB images. The proposed domain adaptation method is also helpful to improve the saliency detection accuracy.

## Acknowledgments

This work is supported by the NSFC 61672089, 61703436, 61572064, 61273274, CELFA; NSFC-61672376, NSFC-U 1803264; NSFC-61603057. We gratefully appreciate the help of Dingxin Yan for image capturing.

## References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1597–1604.

Benjdira, B.; Bazi, Y.; Koubaa, A.; and Ouni, K. 2019. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing* 11(11):1369.

Brown, M., and Süsstrunk, S. 2011. Multi-spectral sift for scene category recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 177–184.

Chen, S.; Tan, X.; Wang, B.; and Hu, X. 2018. Reverse attention for salient object detection. In *European Conference on Computer Vision*, 234–250.

Cheng, M.-M.; Mitra, N.; Huang, X.; Torr, P.; and Hu, S.-M. 2015. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(3):569–582.

Fu, H.; Cao, X.; and Tu, Z. 2013. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing* 22(10):3766–3778.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2018. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(7):1761–1773.

Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; and Li, S. 2013. Salient object detection: A discriminative regional feature integration approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2083–2090.

Jiang, J.; Feng, X.; Liu, F.; Xu, Y.; and Huang, H. 2019. Multi-spectral rgb-nir image classification using double-channel cnn. *IEEE Access* 7:20607–20613.

Li, G., and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5455–5463.

Li, G., and Yu, Y. 2016. Deep contrast learning for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 478–487.

Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; and Shum, H.-Y. 2010. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(2):353–367.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 733–740.

Qu, L.; He, S.; Zhang, J.; Tian, J.; Tang, Y.; and Yang, Q. 2017. Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing* 26(5):2274–2285.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Shen, X., and Wu, Y. 2012. A unified approach to salient object detection via low rank matrix recovery. In *IEEE Conference on Computer Vision and Pattern Recognition*, 853–860.

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

Tu, Z.; Xia, T.; Li, C.; Lu, Y.; and Tang, J. 2019. M3s-nir: Multi-modal multi-scale noise-insensitive ranking for rgb-t saliency detection. In *IEEE Conference on Multimedia Information Processing and Retrieval*, 141–146.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 7167–7176.

Vicente, S.; Kolmogorov, V.; and Rother, C. 2008. Graph cut based image segmentation with connectivity priors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2517–2526.

Wada, K. 2017. pytorch-fcn: PyTorch Implementation of Fully Convolutional Networks. https://github.com/wkentaro/pytorch-fcn.

Wang, Q.; Yan, P.; Yuan, Y.; and Li, X. 2013. Multi-spectral saliency detection. *Pattern Recognition Letters* 34(1):34–41.

Wang, J.; Jiang, H.; Yuan, Z.; Cheng, M.-M.; Hu, X.; and Zheng, N. 2017a. Salient object detection: A discriminative regional feature integration approach. *International Journal of Computer Vision* 123(2):251–268.

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017b. Learning to detect salient objects with image-level supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 136–145.

Wang, X.; You, S.; Li, X.; and Ma, H. 2018. Weakly-supervised semantic segmentation by iteratively mining common object features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1354–1362.

Wang, Q.; Zhu, G.; and Yuan, Y. 2013. Multi-spectral dataset and its application in saliency detection. *Computer Vision and Image Understanding* 117(12):1748–1754.

Wei, Y.; Wen, F.; Zhu, W.; and Sun, J. 2012. Geodesic saliency using background priors. In *European Conference on Computer Vision*, 29–42.

Wu, Z.; Su, L.; and Huang, Q. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yu, H.; Zheng, K.; Fang, J.; Guo, H.; Feng, W.; and Wang, S. 2018. Co-saliency detection within a single image. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhang, P.; Wang, D.; Lu, H.; and Wang, H. 2018. Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. *arXiv preprint arXiv:1802.07957*.

Zhang, F.; Du, B.; and Zhang, L. 2014. Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing* 53(4):2175–2184.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2223–2232.